Engineering System Design for Automated Space Weather Forecast

M H Alomari

PhD

2009

Engineering System Design for Automated Space Weather Forecast

Designing Automatic Software Systems for the Large-Scale Analysis of Solar Data, Knowledge Extraction and the Prediction of Solar Activities Using Machine Learning Techniques

Mohammad Hani Alomari, MSc

A thesis submitted for the degree of

Doctor of Philosophy

School of Computing, Informatics & Media

University of Bradford

2009

September 2009

University of Bradford

To My Uncle,

Dr. Hussein Al-Omari

Abstract

Coronal Mass Ejections (CMEs) and solar flares are energetic events taking place at the Sun that can affect the space weather or the near-Earth environment by the release of vast quantities of electromagnetic radiation and charged particles. Solar active regions are the areas where most flares and CMEs originate. Studying the associations among sunspot groups, flares, filaments, and CMEs is helpful in understanding the possible cause and effect relationships between these events and features. Forecasting space weather in a timely manner is important for protecting technological systems and human life on earth and in space.

The research presented in this thesis introduces novel, fully computerised, machine learning-based decision rules and models that can be used within a system design for automated space weather forecasting. The system design in this work consists of three stages: (1) designing computer tools to find the associations among sunspot groups, flares, filaments, and CMEs (2) applying machine learning algorithms to the associations' datasets and (3) studying the evolution patterns of sunspot groups using time-series methods.

Machine learning algorithms are used to provide computerised learning rules and models that enable the system to provide automated prediction of CMEs, flares, and evolution patterns of sunspot groups. These numerical rules are extracted from the characteristics, associations, and time-series analysis of the available historical solar data. The training of machine learning algorithms is based on data sets created by investigating the associations among sunspots, filaments, flares, and CMEs. Evolution patterns of sunspot areas and McIntosh classifications are analysed using a statistical machine learning method, namely the Hidden Markov Model (HMM).

Acknowledgment

First of all, my strongly thanks for ALLAH the most merciful, without his help and blessing, this thesis would not have progressed nor have seen the light.

I would like to record my gratitude to Dr. Rami Qahwaji for his supervision, advice, and guidance from the very early stage of this research. Above all, he provided me with unflinching encouragement and support in various ways. His most prominently directed guidance and integral view on research, and his mission for providing high-quality work and nothing less has made a deep impression upon me. I appreciate his time and efforts in reviewing my thesis and research papers. I am indebted to him more than he knows. Without Dr. Qahwaji I would not have achieved the objectives of my research.

I would also thank my second supervisor Dr Stan Ipson for his ongoing encouragement, support and advice. I am also grateful for his detailed and precisely pinpointed comments and suggestions while reviewing my thesis and all the published work.

I am also grateful to Dr. Tufan Colak who encouraged and helped me to publish part of my work and gave me many constructive ideas, helpful discussions and appreciated guidance and advice.

And I would also like to give special thanks to the external examiner of my thesis, Prof. Thierry Dudok De Wit (LPCE, CNRS and Université d'Orléans - France) and the internal examiner, Prof Rae Earnshaw for their time, efforts, professionalism and extensive knowledge while reviewing my thesis, giving feedback, and serving as my PhD examiners.

I would like to warmly thank my parents and my wife who have been a constant source of inspiration to me. It is their love and encouragement that has made this long

journey full of joy and has encouraged me to be patient during the PhD period. My special gratitude is due to my brothers and sisters for their loving support.

Finally, I would like to thank all the friends and colleagues at the University of Bradford who constantly provided emotional support and took care of me in many aspects during the three years of studies.

Publications

- M. Al-Omari, R. Qahwaji, T. Colak, and S. Ipson, "Machine Learning-Based Investigation of the Associations between CMEs and Filaments," *SUBMITTED* to Solar Physics, 2009.
- M. Al-Omari, R. Qahwaji, T. Colak, S. Ipson, and C. Balch, "Next-Day Prediction of Sunspots Area and McIntosh Classifications using Hidden Markov Models," in 2009 International Conference on CYBERWORLDS Bradford, UK, 2009.
- R. Qahwaji, T. Colak, M. Al-Omari, O. Ahmed, J. Zraqo, and S. Ipson, "Present and Future Directions in the Automated Prediction of Solar Flares," in Space Weather Workshop: The meeting of Science, Research, Applications, Operations, and Users. Boulder, Colorado: http://www.fin.ucar.edu/UCARVSP/spaceweather/abstract_view.php?recid=1010, 2009.
- R. Qahwaji, M. Al-Omari, T. Colak, and S. Ipson, "Using the Real, Gentle and Modest AdaBoost Learning Algorithms to Investigate the Computerised Associations between Coronal Mass Ejections and Filaments," in Mosharaka International Conference on Communications, Computers and Applications (MIC-CCA 2008), Mosharaka for Researches and Studies, Amman, Jordan, 2008, pp. 37-42.
- R. Qahwaji, M. Al-Omari, T. Colak, and S. Ipson, "Computerised Representation of the Association between Solar Features and Activities using Radial Basis Functions," in IASTED International Conference on Visualization, Imaging and Image Processing (VIIP 2008), ACTA Press, Palma de Mallorca, Spain, 2008, p. 808.

- M. Al-Omari, R. Qahwaji, T. Colak, and S. Ipson, "Support Vector Machines for Automated Knowledge Extraction from Historical Solar Data: A Practical Study on CME Predictions," in 5th International Multi-Conference on Systems, Signals and Devices (IEEE SSD 2008), Amman, Jordan, 2008, pp. 1-6.
- R. Qahwaji, T. Colak, M. Al-Omari, and S. Ipson, "Investigating the Association among Active Regions, Flares and CMEs using Machine Learning," in SOHO 20 Conference: Transient Events on the Sun and in the Heliosphere Ghent, Belgium: http://www.soho20.org/IMG/Session4/SOHO20_S4-P55_qahwaji_Id=120.pdf, 2007.
- R. Qahwaji, T. Colak, and M. Al-Omari, "Large-Scale Numerical Analysis for the Prediction of Flares using Support Vector Machines and Neural Networks," in SOHO 20 Conference: Transient Events on the Sun and in the Heliosphere Ghent, Belgium: http://www.soho20.org/IMG/Session4/SOHO20_S4-P60_qahwaji_Id=122.pdf, 2007.
- R. Qahwaji, T. Colak, M. Al-Omari, and S. Ipson, "Prediction of CMEs Using Automated Machine Learning of CME-Flare Associations," Solar Physics, vol. 248, pp. 471 - 483, 2008.
- M. Al-Omari, R. Qahwaji, T. Colak, and S. Ipson, "Morphological-Based Filtering of Noise: Practical Study on Solar Images," in IEEE International Conference on Signal Processing and Communication Dubai, United Arab Emirates, 2007, pp. 1075-1078.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AdaBoost	Adaptive Boosting algorithm
ASAP	Automated Solar Activity Prediction
BBSO	Big Bear Solar Observatory
CAO	Catania Astrophysical Observatory
CCNN	Cascade-Correlation Neural Network
CME	Coronal Mass Ejection
CPA	Central Position Angle
EIT	Extreme ultraviolet Imaging Telescope
ESA	European Space Agency
EUV	Extreme Ultra-Violet
FAR	False Acceptance Rate
FNR	False Negative Rate
FPR	False Positive Rate
FRR	False Rejection Rate
GHN	Global high-resolution Hydrogen-alpha Network
HAO	High Altitude Observatory
HF	High Frequency
HMM	Hidden Markov Model
HMT	Hit-Miss Transform
HSOS	Huairou Solar Observing Station
HSS	Heidke Skill Score
IMF	Interplanetary Magnetic Field
KSO	Kanzelhohe Solar Observatory
LASCO	Large Angle and Spectrometric Coronagraph
LUF	Lowest Usable Frequency

MDI	Michelson Doppler Imager
MLSO	Mauna Loa Solar Observatory
MPA	Measurement Position Angle
MUF	Maximum Usable Frequency
NASA	National Aeronautics and Space Administration
NGDC	National Geophysical Data Centre
NN	Neural Network
NOAA	National Oceanic and Atmospheric Administration
RBFN	Radial Basis Function Network
ROC	Receiver Operating Characteristic
SID	Sudden Ionospheric Disturbance
SMM	Solar Maximum Mission
SOHO	SOlar and Heliospheric Observatory
SOLWIND	SOLar WIND coronagraph
SVM	Support Vector Machine
SWF	Short Wave Fade
SWPC	Space Weather Prediction Centre
SXT	Soft X-ray Telescope
TNR	True Negative Rate
TPR	True Positive Rate
YNAO	YunNan Astronomical Observatory

CHAPTER ONE

1 INTRODUCTION

1.1 Background

Generally, the importance of space weather is increasing as more human activities take place in space and as we rely more and more on communications and power systems. Space weather is defined as "the time-variable conditions in the space environment that may affect space-borne or ground-based technological systems and, in the worst case, endanger human health or life." (Koskinen et al., 2001).

Solar flares and Coronal Mass Ejections (CMEs) are dramatic indicators at the Sun of imminent adverse space weather. Flares and CMEs are two types of solar eruptions that can spew vast quantities of radiation and charged particles into space (Lenz, 2004). The Earth environment and geomagnetic activity are affected by the ionized solar plasma, also known as the solar wind. The solar wind, affected by solar activity, flows outward from the sun, carrying with it the magnetic field of Sun, to form the heliosphere (Pick et al., 2001). The resulting Interplanetary Magnetic Field (IMF) creates storms by injecting plasma into the Earth's magnetosphere (Yurchyshyn et al., 2003); (Yevlashin and Maltsev, 2003). Geomagnetic storms are correlated with CMEs (Wilson and Hildner, 1984) and predicting CMEs can be useful to forecast space weather (Webb, 2000). Major solar flares can also seriously disrupt the ionosphere and in order to guarantee that humans can work safely and effectively in space, the forecasting of strong solar flares is also important (Kurokawa, 2002).

Previous researches on CMEs, such as Munro (1979), Poland et al. (1981), Moon et al. (2002), Jing et al. (2003), Zhou et al. (2003), Yashiro et al. (2005) and Yashiro et al. (2006), reported that most of the CME events are associated somehow with eruptive filaments/prominences and/or solar flares. The studies Severny (1965), Warwick (1966), Sakurai (1970), and McIntosh (1990) on solar flares showed that they are mostly related to sunspots and active regions. Sunspots are part of active regions, and their local behaviour is used for the forecast of solar activity (Hathaway et al., 1994).

The first CME in the astronomical literature was reported in 1860 and re-discovered in the 70's (Briand, 2003). CMEs are huge bubbles of gas that are ejected as a sporadic expulsion of mass from the solar corona to the interplanetary medium. In September 1859, Richard Carrington and Richard Hodgson independently recorded the first solar flare (Carrington, 1859, Tassoul et al., 2005). Flares are violent explosions that occur due to the sudden release of magnetic energy that has been building in the solar atmosphere. Solar flares and Coronal Mass Ejections (CMEs) are the most dramatic solar events affecting the terrestrial environment (Pick et al., 2001).

For years, solar flares were thought to be responsible for large perturbations in the solar wind and geomagnetic environment. However, space based chronographs have made us aware of CMEs (Tousey, 1973). A pioneering and controversial work (Gosling, 1995) argued that CMEs, not flares, were the critical element for large geomagnetic storms, interplanetary shocks, and major solar energetic particle events. This contradicted the findings of Lin and Hudson (1976) that flare accelerated particles in big flares provides the energy for all the activities that followed such as CMEs and large energetic particles events. Since then there have been many studies aiming to find out how CMEs are initiated and triggered.

Currently, five major CME eruption models exist: the Thermal Blast Model, the Dynamo Model, the Mass Loading Model, the Tether Release Model and the Tether Straining Model (Klimchuk, 2001, Low, 1999b, Low, 2001a, Low, 2001b). The last three are storage and release type models, where a slow build-up of magnetic stress occurs before eruption begins (Aschwanden, 2004). The model that is considered in the present thesis is the mass loading model which can explain some cases of CMEs that are associated with filaments/prominences. The mass loading process during the pre-eruption phase of a CME can be manifested in the form of a growing quiescent or eruptive filament. Mass loading can be associated with prominences, which are extremely dense and of chromospheric temperature, contained in a compact volume. Prominences are thought to play a major role in CME eruptions because of their simultaneous appearance, according to the observations reported in (Low, 1996, Low, 1999a). A crucial criterion for CME eruptions is the mass of the prominence and its role in the storage of magnetic energy (Low et al., 2003, Zhang and Low, 2004).

1.2 Space Weather: Causes and Effects

The best way to understand space weather causes and effects is to study how the Sun can affect the space environment. On average, the Sun is 149.6 million kilometres away from the Earth. The Sun can be divided into two main parts: the solar interior and the outer surface. As shown in Figure 1.1, the solar interior is composed of the core, the radiative zone, and the convective zone. The Sun's outer surface includes the photosphere, the chromosphere and the corona. All the heat and light from the Sun that we detect is produced originally in the Sun's core through complex nuclear reactions.

The space weather disturbances are caused mainly by the Sun, through the different types of solar activities shown in Figure 1.2. These activities affect the space and ground-based systems according to the time scale depicted in Figure 1.3.

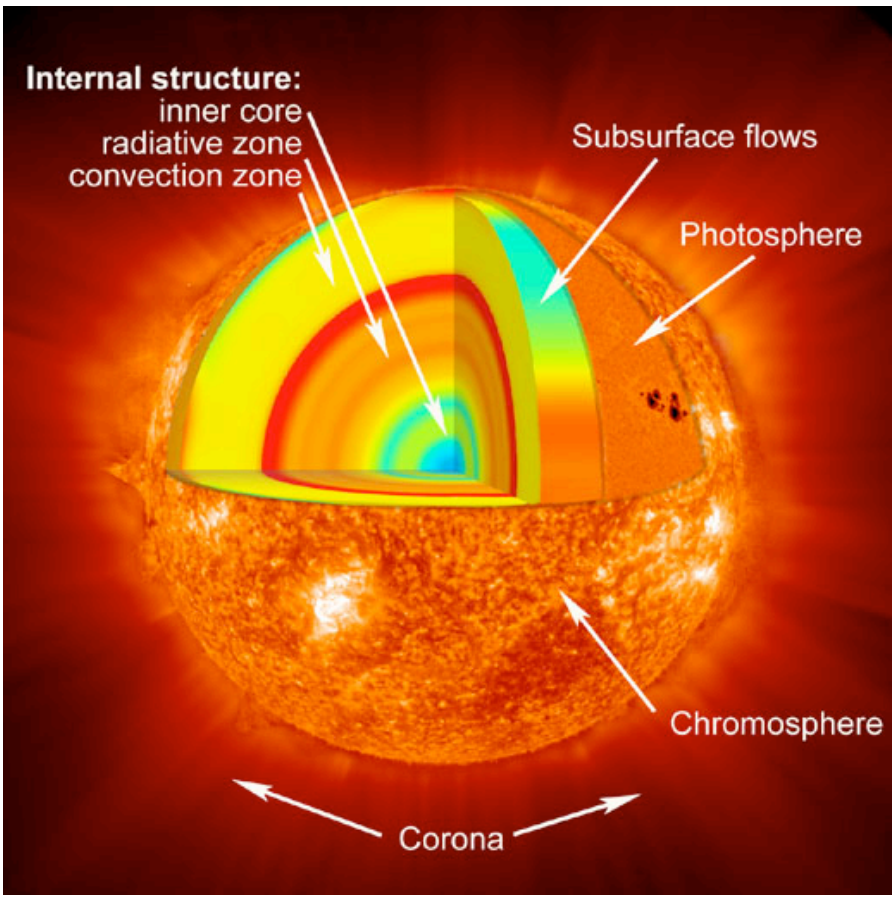


Figure 1.1 Layers of the Sun. Image courtesy NASA.

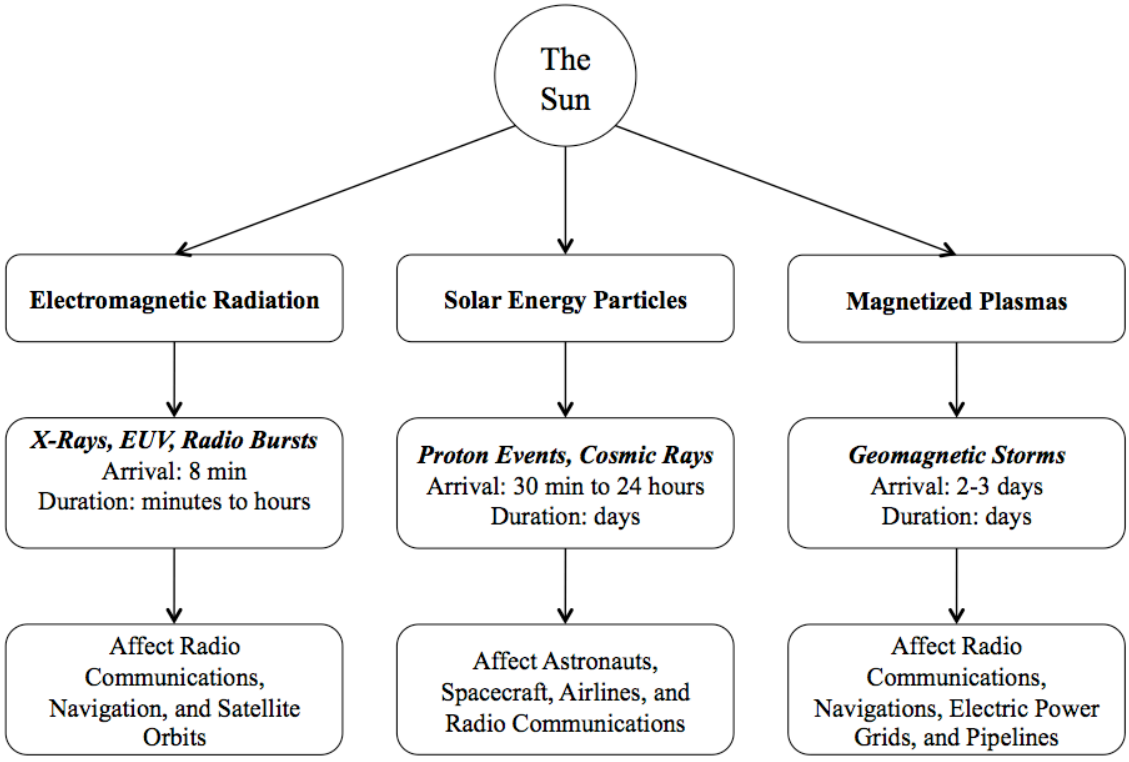


Figure 1.2 Types of space weather causes and effects.

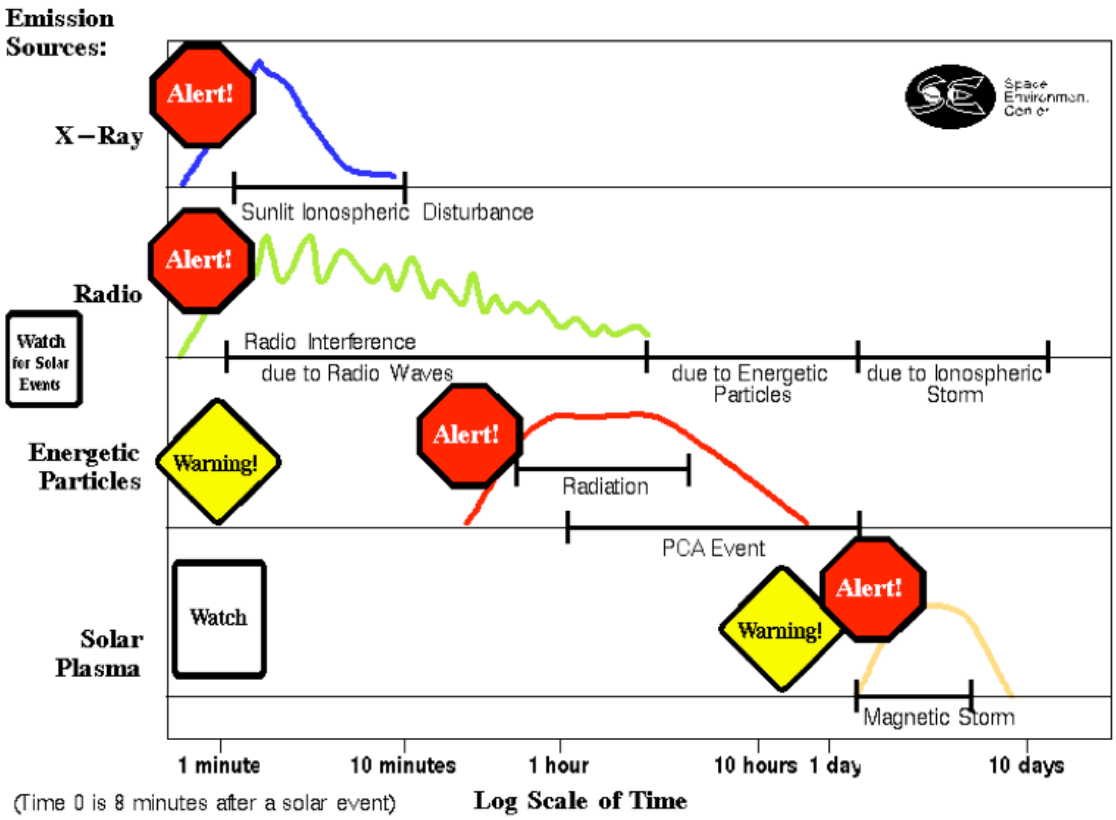


Figure 1.3 Time scale of emission sources and solar effects. Image courtesy the media and graphics center at NOAA Space Environment Center.

Drivers of space weather can be grouped based on causes and effects shown in Figure 1.2 as follows:

- *Electromagnetic radiation produced by significant variations of the invisible solar photons over the solar cycle.* X-ray and Extreme Ultra-Violet (EUV) flux from flares, travelling at the speed of light, interacts with the ionosphere, producing wide spread blackout conditions for High Frequency (HF) radio communications (Davies, 1989). The most important Sudden Ionospheric Disturbance (SID), affecting communication systems, is the Short Wave Fade (SWF). The transmitter-receiver satellite systems use propagation frequencies in the usable frequency range between the Maximum Usable Frequency (MUF) and the Lowest Usable Frequency (LUF) as shown in Figure 1.4. During significant X-Ray events, the ionization and absorption of the

ionosphere's lowest portion are enhanced which raises the LUF and causes a SWF. This means that short HF waves would be absorbed or reflected by the particles in ionosphere's lowest layer causing a complete blackout of radio communications (Figure 1.5).

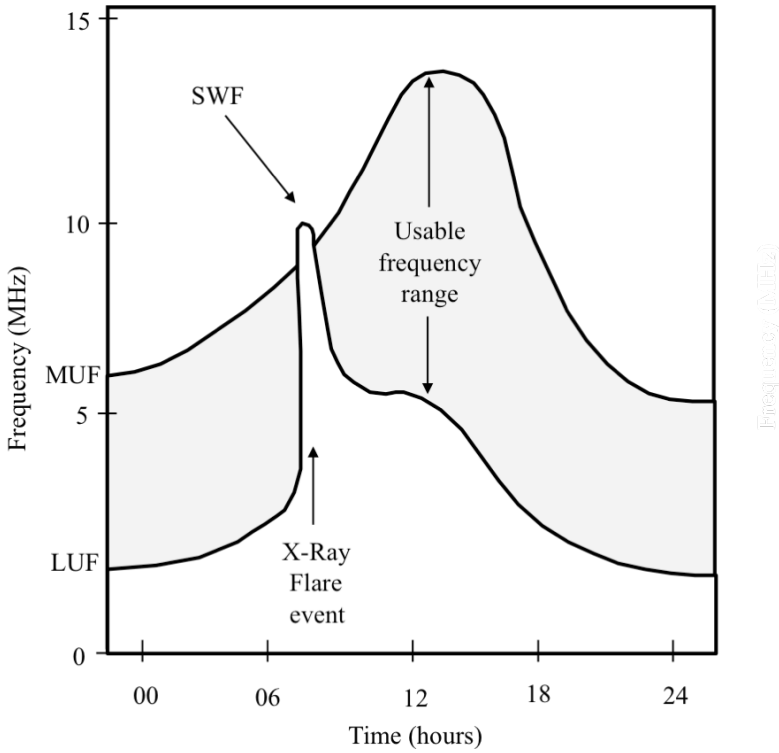


Figure 1.4 Short wave fade because of a flare X-Ray event. Image courtesy of the media and graphics center at NOAA Space Environment Center [modified].

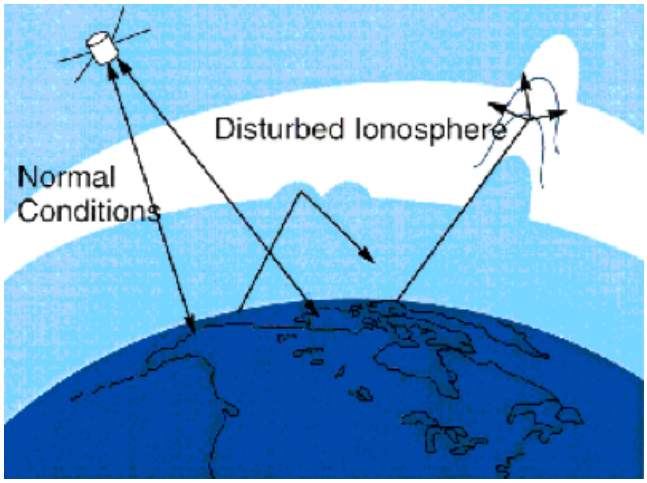


Figure 1.5 Blackout of radio communications (short wave fade). Image courtesy the media and graphics center at NOAA Space Environment Center.

- *High energy solar particles (Proton events, sometimes travelling at close to the speed of light).* Some forms of solar activity are closely associated with the production of high energy particles which can arrive at the Earth within 15 minutes after the electromagnetic events, resulting in a serious radiation hazard to astronauts in space missions, increased incidence of spacecraft anomalies, and HF communication outages in the polar regions (Balch, 2008).
- *Magnetized plasmas of the solar wind (Geomagnetic storms).* Another space weather effect can result if the outwardly propagating CME arrives at Earth and is able to effectively transfer energy into the Earth's magnetosphere, leading to a geomagnetic storm (Gonzalez and Tsurutani, 1987). Geomagnetic storms are known to affect electrical power grids, global satellite navigation systems, satellite and HF radio communication systems, and frictional drag affecting low-Earth orbiting satellites. During large geomagnetic storms, satellites and their electronics can be damaged. Power grids, for example, can be overloaded because of the large electric fields and currents induced by the currents flowing in the ionosphere. As shown in Figure 1.6, ionosphere currents of up to millions of Amperes may induce large flows in power lines, pipelines, or in large conductors like seawater and the rocks in the Earth's crust. In addition systems that are based on electronic navigation can face potential danger due to errors caused by shifting of radio waves. Surveyors and geologists use Earth's magnetic field in their mapping work and in searching for oil, gas or mineral deposits. So, their measurements can be affected during geomagnetic storms.

In summary, space weather hazards can affect both space and the Earth as depicted in Figure 1.7.

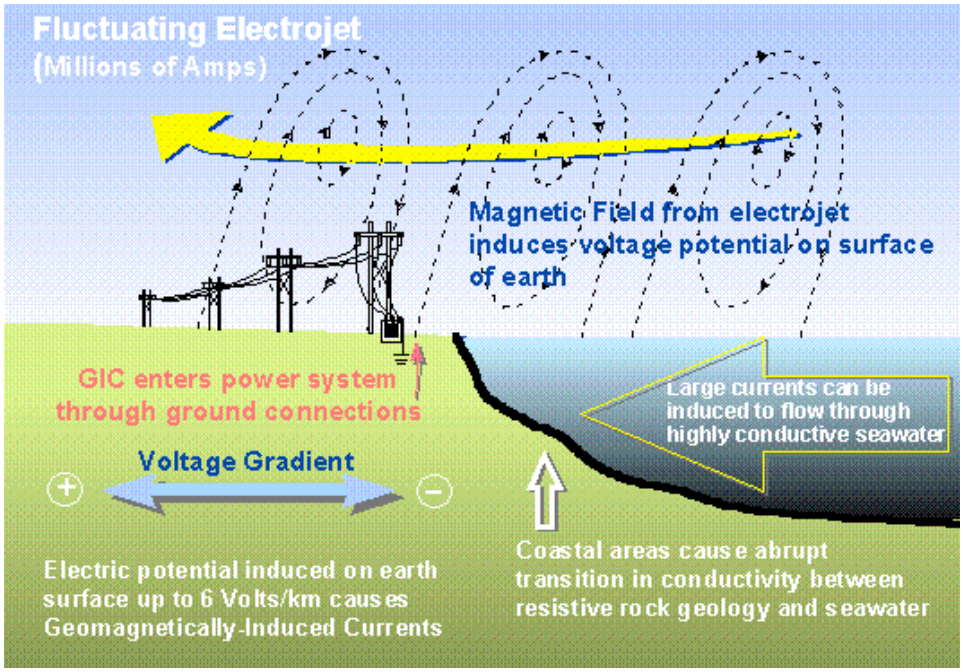


Figure 1.6 Geomagnetic effects on electric power grids. Image courtesy John G. Kappenman, Minnesota Power, Duluth, Minnesota.

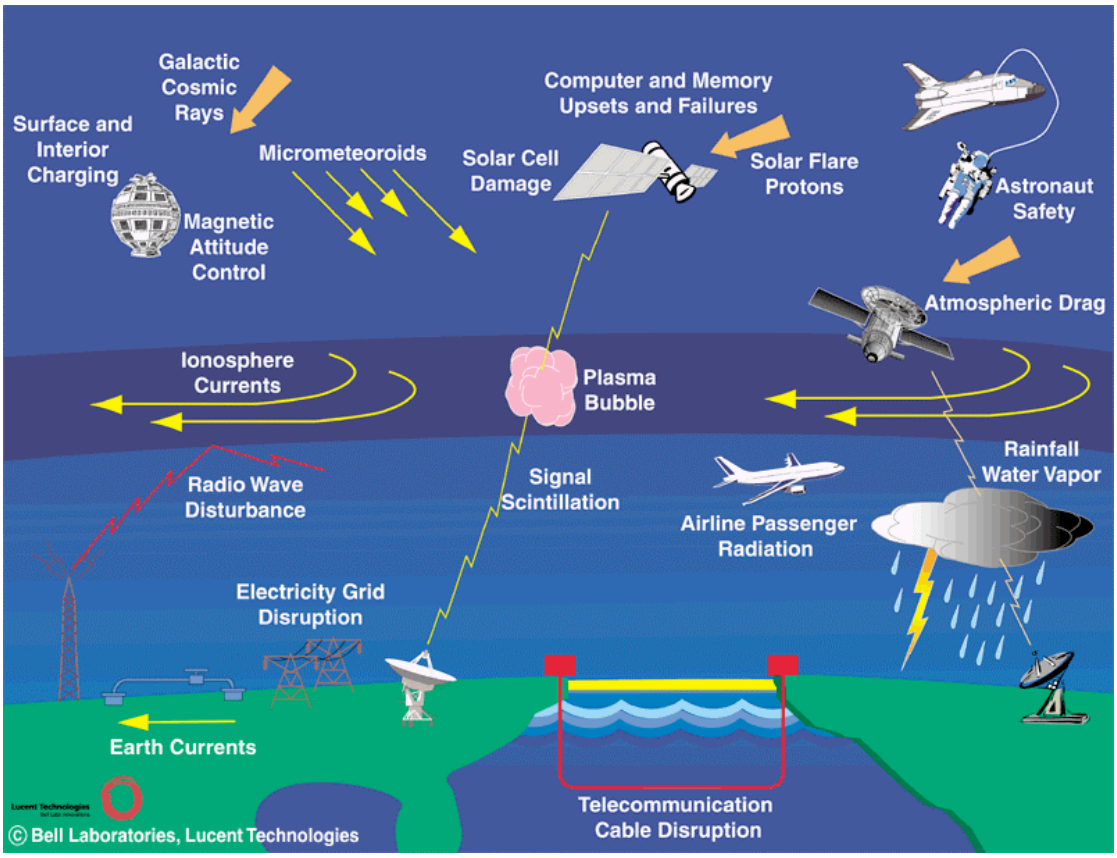


Figure 1.7 Space Weather Hazards, Image courtesy Lou J. Lanzerotti, Bell Laboratories, Lucent Technologies, Inc.

1.3 Motivation

It is shown in the previous section how space weather disturbances can be directly or indirectly hazardous to both humans and systems. In the case of a significant solar event occurring, predicting space weather conditions could help in taking proper actions to minimize any possible damage. For example, navigators can use backup navigation systems, power providers can protect their systems to avoid power outages, and surveyors and geologists can reschedule their activities.

With the rapid development and the wide use of technology systems that can be affected by space weather conditions, it is more important to forecast space weather. For space weather alerts and warnings to be helpful in our life, automated real-time (or near real-time) prediction systems are needed.

Because of the solar observations provided by many space missions and ground-based observatories, the recent space missions (Hinode and STEREO), and the expected space missions (EDO) data volumes will increase 1000 to 10,000 times and data has already started to pile-up. Hence, the design of automated space imaging systems is becoming more important than ever and it is necessary to combine the available data with reliable data processing computer systems.

In this research the focus is on the design, development, validation and evaluation of automated space weather prediction technologies that, in the future, can be integrated under one forecasting system. The approach in this work is on defining some of the space weather research needs first, then proceeding with the design synthesis and validations while considering the complete problem of automated space weather forecasting. The engineering system design process followed in this research can be depicted as shown in Figure 1.8. It is aimed in this process to make efficient use of the available solar data, computer tools, solar physics models, and mathematical algorithms.

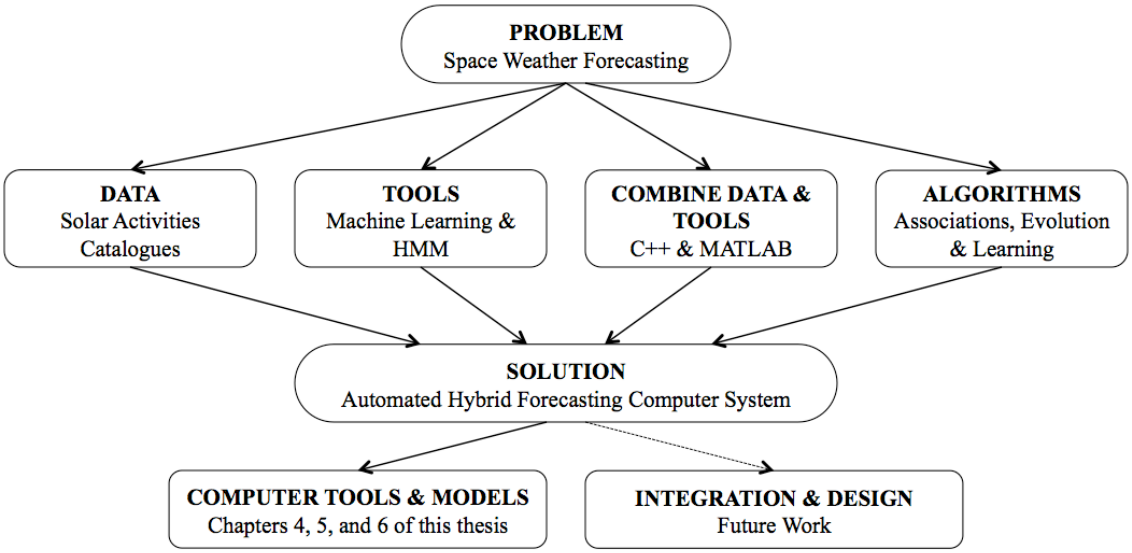


Figure 1.8 Research Work Organization.

Oliver et al. (1997) described a model that decomposes the process of systems engineering into two processes: a management process and a technical process. The management process aims to organize the technical efforts in the system design while the technical process is needed to review the available information and data sources and define the possible efficiency measures which enable performing the validation analysis and building the complete system. Hence, we can say that the work presented in this thesis will be considering the technical process and the design of many prediction models that can be used in the future to provide a complete forecasting system.

1.4 Research Aims and Objectives

The main objective of this research is to provide computerized decision rules and models for the purpose of automated space weather forecasting. The outcomes of this research will be mainly the design of some new technologies that can be developed in the future within the context of a real-time prediction system. The inputs to such system could be several of the real-time solar images that are available online and provided by many satellite and ground-based solar observatories.

Initially the aim is to analyse the associations between CMEs, flares, and filaments by processing the available data in catalogues. The association datasets will

then be processed using machine learning algorithms to provide computerised learning rules that can be used in a reliable CME predictions system. For the purpose of modelling the sunspot evolution patterns, historical sunspot data will be analysed using Hidden Markov Models (HMMs). It is aimed by such analysis to provide computerized models that can be used to predict the McIntosh classifications and the sunspot areas for the sunspot group under investigation within the next 24 hours.

As a summary, the objectives of this research are:

- To implement a large-scale numerical analysis investigating the associations among sunspot groups, filaments/prominences, solar flares and CMEs covering most of the available solar data in solar cycle 23.
- To implement, validate, and make use of some of the scientific initiation models in the field of solar physics using computerised algorithms.
- To compare machine learning algorithms for the efficient knowledge extraction and to represent the evolution patterns of sunspot groups, flaring activity, and CME eruptions using computerized learning rules and models.
- To propose a fully automated computer platform, based on the learning rules, that could provide short-term predictions for the possible active region flaring and CME eruptions.

1.5 Original Contributions

The main original contributions presented in this thesis can be summarised as follows:

- A computer tool is designed to process the available solar data in catalogues and analyse them to find the associations between solar features and events. Different association levels and algorithms were

implemented to extract the associations among sunspot groups, filaments, solar flares, and CMEs covering most of solar cycle 23.

- Machine learning algorithms were used for the first time within the context of automated CME predictions. The system design implements different machine learning algorithms which have been optimised to create three sets of computerised decision rules representing the associations between: (1) CMEs and flares (2) CMEs and filaments and (3) CMEs and sunspot-associated flares.
- Hidden Markov Models (HMMs) and the Baum-Welch algorithm are used for the first time to provide computerised models that best fit the evolution patterns of sunspot groups. The time-series analysis of the historical sunspot data enabled the model to provide predictions for the sunspot area and McIntosh class for the next 24 hours.
- For the purpose of solar flare predictions, a future plan is provided in an attempt to model the relationship between flares and sunspots using HMMs. In addition, the associations between sunspots and flares were analysed using machine learning algorithms to improve the previous attempts of flare predictions reported in the literature.

1.6 Outline of the Thesis

This thesis is organized as follows:

- Chapter 2 provides an extended literature review for recent research on associations and predictions of solar activities.
- Chapter 3 explores the available sources of solar data that can be used in the research presented in this thesis. It also describes the association principles and discusses different levels of associations between sunspot groups, filaments/prominences, solar flares and CMEs.

- Chapter 4 provides a practical implementation and an evaluation for the proposed forecasting systems using machine learning algorithms. It compares the performances of many learning algorithms: Cascade-Correlation Neural Networks (CCNNs), Support Vector Machines (SVMs), Radial Basis Function Networks (RBFNs) and the Adaptive Boosting (AdaBoost) algorithm.
- Chapter 5 studies the evolution of sunspot groups using Hidden Markov Models (HMMs). In addition, it describes the development of a model that can be used to predict the McIntosh class and the sunspot area in an attempt to enhance the previous work on solar flare prediction.
- Concluding remarks and recommendations for future work are presented in Chapter 7. In addition, Chapter 7 introduces the hybrid space weather forecasting system which is a practical implementation integrating the technologies developed in this work.

CHAPTER TWO

2 LITERATURE REVIEW

2.1 CMEs: Cause and Effect

Reading the literature the cause and effect relations between CMEs and other solar events are not clear. Webb et al. (1998) reported a case study of the associations between CMEs, magnetic clouds, and geomagnetic storms and found that CMEs are the real link between solar eruptions and space weather activities affecting the Earth. The assumption of a direct CME-flare relationship has driven most of the solar flare myth controversy (Cliver and Hudson, 2002). On the other hand, there are also many studies on the solar origin of CMEs such as Zhang and Wang (2001) which found associations between large CMEs and filament eruptions. By overlaying EIT and LASCO C2 images for a selected active region with the same spatial scales, it was found that each CME component corresponds to a filament eruption and a flare. Recently, new findings in Robbrecht et al. (2009) showed that the initiation of a CME does not need to be associated with clear on-disk activities. These authors analysed a large-scale front side CME observed by the SECCHI instruments onboard the STEREO mission and found that this CME has neither Hα disk signatures nor observable filaments.

Zhang et al. (2001) measured the CMEs initial evolution in the low corona and then explored the possible causes of CME initiation and acceleration in connection with flares. The kinematical evolution of CMEs is described in a three-phase scenario: the initiation phase, the impulsive acceleration phase, and the propagation phase.

MacQueen and Fisher (1983) and Sheeley et al. (1999) reported that the key observation is the acceleration profile of the CME (or filament) during the flare. That is; if no flare or only a weak flare occurs, then we would have the slowly-accelerating eruptive filament events but if a flare occurs, then an additional acceleration process might act on the CME. The changes associated with the magnetic topology for the X1.2 flare that occurred on 30 September 2000 and was not associated with a CME were studied in Green et al. (2002). It was noted that the flare resulted from the interaction of two pre-existing loops low in the corona which produced a confined flare. Hillaris et al. (2006) investigated the intensity, impulsiveness and energetics of solar flares with and without associated CMEs for the period between 1998 and 2000. It was found that flares that were not associated with any CME and associated with type II metric bursts were the most impulsive, having the shortest duration.

In an attempt, in Mouradian et al. (1995), to better understand the associations between eruptive filaments/prominences and CMEs, the sudden disappearance (DB) of quiescent filaments/prominences was divided into two classes: dynamic and thermal disappearances. The dynamic DB is considered to consist of an expansion and ejection of prominence plasma into the corona due to changes in the underlying magnetic field structure, such as the emergence of new magnetic flux. On the other hand, the thermal DB is considered to consist of the disappearance of prominences in H-alpha line due to an energy increase. This study showed that dynamic DBs are associated with CMEs, whereas thermal DBs are just local disturbances at the lower corona. Pojoga and Huang (2003) made a similar study on the sudden disappearances of prominences/filaments for the period between January and April 2000 and studied their correlation with CMEs. Similarly to Mouradian et al. (1995), it was concluded that 70% of the eruptive filaments are associated with CMEs, while the correlation is weaker for the quasi-eruptive and vanishing filaments. In this work, the term “vanishing” is used when

referring to the thermal disappearances of prominences/filaments. This contradicts the findings of Yang and Wang (2002) where the association between CMEs and eruptive filament/prominence is found to be as low as 30%. Yang and Wang (2002) carried out a statistical study of 431 filament/prominence disappearances compiled from BBSO H-alpha images and observed between January 1997 and June 1999. However, they stated that they didn't make a distinction between thermal filament disappearances and filament eruptions. In addition, filament disappearances on disk might be associated with very weak halo CMEs which are difficult to detect.

Some researchers have examined the relationship between CMEs and filament disappearances. A statistical study of filament disappearances between 1999 and 2002 is presented in Jing et al. (2003). They studied 79 events where the H-alpha or EIT/LASCO observations during the filament disappearance are available. They found that 63% of these events are associated with CMEs (excluding 6 events with no LASCO data) and it is more likely for filament eruptions to be associated with CMEs than flares. The flares association was found to be 91% with active region filaments and 20% with quiescent filaments. CMEs associations were found to be about 73% with active region filaments and 55% with quiescent filament eruptions.

2.2 Large-Scale Analysis

To draw more accurate and meaningful conclusions, some researchers examined the correlations between CMEs and solar surface activities for large numbers of solar events. For example, Moon et al. (2002) analysed 3217 CME events observed by SOHO/LASCO from 1996 to 2000 and made a statistical study of their associations with solar flares using GOES X-ray images and eruptive filaments using H-alpha images from BBSO. They found that the CMEs that are associated with flares had larger velocities. Moon et al. (2003) surveyed all CMEs observed by SOHO/LASCO for the period from 1997 to 2001 and selected 197 front-side halo CMEs. They concluded that

88% of the halo CMEs were associated with flares and more than 94% were associated with eruptive prominences/filaments, while 79% of the CMEs were initiated from active regions.

Most of the large-scale association studies reported in the literature concentrate on the relationship between flares and CMEs. In Andrews (2003), 311 M and X-class flares, which occurred during the years 1996 to 1999, were investigated to find their associated CME candidates. The SOHO/LASCO CME data were used in this study. Online catalogues were used to search for CME candidates for the 229 flares with good LASCO data coverage. It was found that about 40% of the M-class flares do not have associated CMEs and the probability of finding a CME candidate for the association does not depend on the location of the flare.

CME-associated flares distributions were analysed statistically in Shrivastava and Singh (2005). They studied the latitudinal locations for the flares in the northern and southern hemispheres for the period 1986 to 2003. It was found that CME-associated flares are equally distributed in the northern and southern hemispheres. There is also some research on the intensity of the solar flares and CMEs. Yashiro et al. (2005) examined the CME visibility (detection efficiency) for 1301 X-ray flare events above C3 level (49 X-class, 610 M-class, and 642 C-class flares) from 1996 to 2001. It was assumed that all CMEs associated with limb flares are detectable by LASCO. Based on a statistical study of the properties of the flare-associated CMEs and a comparison with flare size and longitude it was found that the CME association rate increased with the flare size from 20% for C-class flares to 100% for huge X-class flares. It was also concluded that CMEs associated with disk C-class flares were slower and narrower than those of CMEs associated with X-class flares.

A discussion of the associations of CMEs with flare properties is presented in Yashiro et al. (2006). Properties such as peak X-ray intensity, total X-ray intensity, and

the decay time for 1540 X-ray flares (M-class and above, including 50 huge flares above X1.8) were analyzed. It was found that CMEs associated with flares above X1.8 have CMEs association rate of 98% compared with only 40% for CMEs associated with flares between M1.0 and M1.7. Also it was concluded that a definite association between CMEs and flares exists if the decay time of the flare exceeds 90 min.

2.3 Selected Events Studies

In Munro et al. (1979), 75 major CMEs observed with the white light coronagraph on Skylab in the period between 1973 and 1974 were surveyed to study their association with other solar activities. It was found that 75% of the CMEs observed were associated with other forms of solar activity, 40% of the CMEs were associated with H-alpha flares, and 50% of the CMEs were associated with eruptive prominences. Another study based on white light coronal images is the work reported by Poland et al. (1981). The Naval Research Laboratory's Earth orbiting coronagraph (SOLWIND) was used for observing CMEs between 1971 and 1974. It was concluded that 50% of the observed CMEs were associated with definite or probable flares or eruptive prominences.

Some researchers concentrated their analyses on the Solar Maximum Mission (SMM) data. Webb and Hundhausen (1987) considered 58 CMEs observed in 1980 by the High Altitude Observatory (HAO) Coronagraph/Polarimeter on the SMM satellite and compared them with other forms of solar activity (eruptive prominences, H-alpha flares, soft X-ray events, and metric type II and IV radio bursts). They found that 66% of the CMEs were associated with these solar activities. Out of these CMEs, 68% were found to be associated with eruptive prominences, 37% were associated with H-alpha flares, 76% were associated with X-ray events and 32% were associated with Radio II, or IV events. Another study on the SMM data is the research work reported in St. Cyr and Webb (1991) where 73 CMEs from 1984 to 1986 are considered. They found that

76% of the CMEs were associated with eruptive prominences, 26% were associated with H-alpha flares and 74% with X-ray events. Srivastava et al. (1997) studied 14 CMEs observed by SMM during the period from March to September 1980 and concluded that strong association existed between CMEs and coronal holes, eruptive prominences and current sheets.

St. Cyr et al. (1999) examined 141 CMEs using the Mark III (MK3) K coronameter at the Mauna Loa Solar Observatory (MLSO) between 1980 and 1989. It was found that 55% of the CMEs were associated with active regions and 82% were associated with eruptive prominences.

In Gilbert et al. (2000) 54 H-alpha events from February 1996 to June 1998 were surveyed. The associations of eruptive prominences and active prominences with CMEs were studied using H-alpha observations that were obtained from MLSO. It was found that 92% of the eruptive prominences and 46% of the active prominences were associated with CMEs.

The sources of 32 CMEs observed between January 1996 and May 1998 were studied and compared using MDI and several H-alpha images in Subramanian and Dere (2001). It was found that 41% of the CMEs were associated with active regions without prominence eruptions, 44% were associated with eruptive prominences embedded in active regions, and 15% were associated with eruptive prominences that took place outside active regions.

Hori and Culhane (2002) used microwave images from the Nobeyama Radioheliograph, to examine 50 prominence eruptions near solar maximum between 1999 and 2000 and showed that 92% of the prominence eruptions were associated with CMEs.

In the same manner Jing et al. (2004) performed a statistical study of 106 filament eruptions detected using H-alpha images from BBSO between 1999 and 2003

and their relations to flares and CMEs. According to their study 56% of the filament eruptions were associated with CMEs. They also classified filament eruptions as active region filament eruptions and quiescent filament eruptions and found that active region filament eruptions had higher flare association (95%) compared to quiescent filament eruptions (27%). They found that quiescent filament eruptions were mostly accompanied by CMEs rather than flares. The prominence eruptions were classified by Gopalswamy et al. (2003) as radial and traverse depending on the direction of their movement (radial or horizontal). The associations with CMEs were investigated as well. Microwave images from the Nobeyama Radioheliograph of 186 prominence eruptions from 1 January 1996 to 31 December 2001, covering the minimum and maximum periods of the current solar cycle 23 were used. It was found that 82% of the prominence eruptions were dominantly radial events while only 18% were traverse events and 72% of the prominence eruptions were found to be clearly associated with CMEs. They also found that 83% of the radial events were associated with CMEs

Active regions were also investigated when studying the CME-flare associations. In Green et al. (2001) flares were examined in nine active regions with CME signatures. It was indicated that the energy released by flaring from the magnetic field of an active region was greater mainly before the CME launch. In Akiyama et al. (2006) the CMEs association rate for two active region flares was examined. Active region 10039 produced three X- and eight M-class flares and the CME-flare association rate was found to be 72%. The CMEs from this active region had an average speed of 1195 km/s speed and an average width of 246°. On the other hand, active region 10044 produced 9 M-class flares, the association rate was found to be 13%, and CMEs from this region had an average speed of 282 km/s speed and an average width of 12°.

2.4 Machine Learning

Despite the recent advances in solar imaging, machine learning and data mining have not been widely applied to solar data. Recently, several learning algorithms (i.e. Neural Networks (NNs), Support Vector Machines (SVMs) and Radial Basis Function Neural Networks (RBFNNs)) were optimized and then compared for the automated short-term prediction of solar flares (Qahwaji and Colak, 2007). The machine learning-based system, reported by Qahwaji and Colak (2007), accepted two sets of inputs: the McIntosh classification of sunspot groups and real-time estimation of the solar cycle.

Borda et al. (2002) described a method for the automatic detection of solar flares using the Multi-Layer Perceptron (MLP) with backpropagation training rule, where a supervised learning technique that required a large number of iterations was used. Qu et al. (2004) proposed a method for automatic solar flare tracking using SVM. The classification performance for features extracted from solar flares was compared by Qu et al. (2003), where each flare was represented using nine features. However, these features provided no information about the position, size and verification of solar flares. Qahwaji and Colak (2006a) used a NN after image segmentation to verify the regions of interest, which were solar filaments.

Qahwaji and Colak (2007) and Colak and Qahwaji (2007a) proposed an Automated Solar Activity Prediction (ASAP) system to predict flares based on the automatic detection and classification of sunspot groups analysing their complexities and areas as they appear on the solar images. ASAP's predictions are generated by analysing the recent MDI images using a combination of advanced imaging and machine learning algorithms.

One of the parametric methods of statistical machine learning is the Hidden Markov Model (HMM). The theory of HMMs was first implemented in the 1970s for applications in the field of speech processing (Bahl and Jelinek, 1975, Baker, 1975,

Jelinek, 1969, Jelinek et al., 1975). Generally, HMMs are known for their applications in pattern recognition. For example, Gellert and Vintan (2006) used HMMs to model the movement sequences of a person inside an office building. They found that their model could be used for the prediction of the next movement of a person, within the building, with an accuracy of 92%.

2.5 Summary and Conclusions

From previous research it can be shown that there is a degree of association between CMEs on one hand and flares and erupting filaments/prominences on the other hand. The exact degree of association is not clear though because most of the available studies were carried out on only a few years of data or on limited cases and using physics-based modelling. In some cases, contradicting findings were reported. For example, not all researchers agree that strong relations exist between CMEs and filament/prominence eruptions. In Yang and Wang (2002) it was found that the association rate was about 30% only. On the other hand, it was reported in Zhou et al. (2003) that more than 94% of the considered CMEs were associated with eruptive prominences/filaments. It is one of the aims of this thesis to study such contradictions and investigate the connection between CMEs, filaments, and solar flares. And it is believed that this study will enable the efficient prediction of space weather.

Overall, it is clear that there have been limited studies with large-scale processing and analysis for years of solar data to explore the associations between CMEs and other solar activities. This is because automated data processing algorithms are not widely developed and used. Also it has been noted that data mining and machine learning have not been implemented before to verify this association and to represent it using computer-based learning rules that could be used to extract knowledge and provide predictions by analysing recent data in real-time mode. In addition, it was concluded that the key parameters for properly forecasting CME are the photospheric

and the coronal magnetic fields, which are not properly observed. The horizontal component of the photospheric field is the only observable component and the coronal field is not measured at all. It is intended that the research work presented in the chapters to follow in this thesis will tackle these issues by advancing the state of the art technologies that can be integrated under an engineering system design for space weather forecasting.

In the next chapter, many types of solar data will be presented. Most of the available data in solar cycle 23 will be used in Chapter 4 to verify different levels of associations between (1) CMEs and flares (2) CMEs and filaments (3) sunspots and flares and (4) between CMEs and sunspot-associated flares.

CHAPTER THREE

3 SOLAR DATA

3.1 Introduction

This chapter introduces different types of solar data that are used in the research work presented in this thesis. Some types of solar data are used to find the associations among sunspots, filaments, flares, and CMEs (Chapter 4) and other types of data are used to verify these associations. In addition, some of the data types are found to be very useful to study the evolution patterns of sunspot groups (Chapter 6).

In general, there are two main types of publically available solar data: solar images and data catalogues. Examples of sources and types of solar images are given in Section 3.2 and data catalogues for sunspot groups, filaments/prominences, solar flares, and CMEs are described in Section 3.3. Some conclusions drawn about these data sources are discussed in Section 3.4.

3.2 Solar Images

Solar images are available from many satellites and ground-based observatories. However, because satellites are above the clouds and outside Earth's atmosphere, their images are better than those from the ground-based telescopes.

3.2.1 Satellites

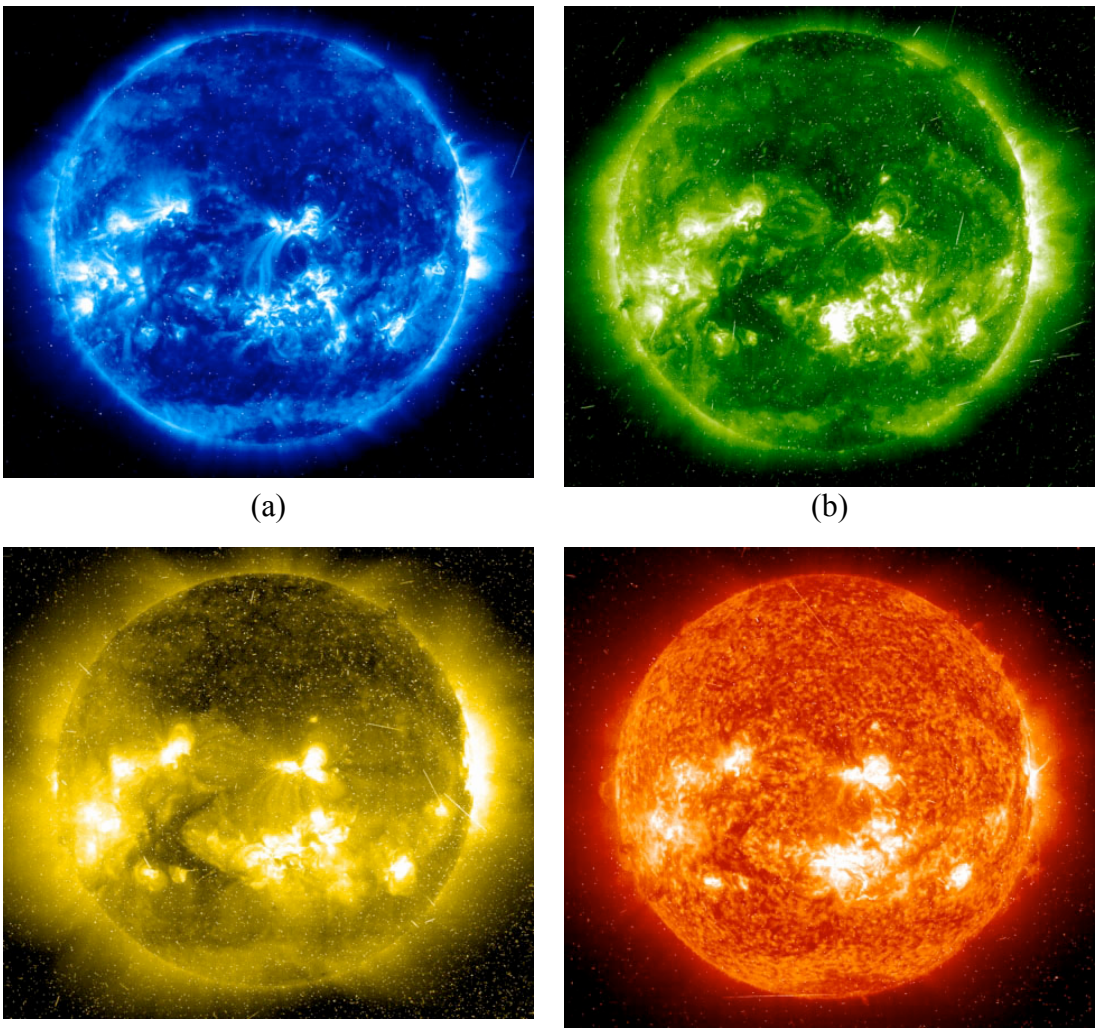
There are many solar observatory satellites such as NASA's Solar and Heliospheric Observatory (SOHO) and Yohkoh. SOHO is a cooperative project

between the European Space Agency (ESA) and the National Aeronautics and Space Administration (NASA) of the USA to study the Sun. Yohkoh ("Sunbeam") is a cooperative mission of Japan, the USA, and the UK and observes the solar atmosphere using radiation in the X-ray region of the spectrum.

This subsection will focus on SOHO because of its importance in the current research. SOHO stays between the Earth and Sun at all times in an orbit which is approximately 1.5 million kilometres away from Earth. There are three main instruments on SOHO providing different types of solar images, including the Extreme ultraviolet Imaging Telescope (EIT), the Large Angle and Spectrometric Coronagraph (LASCO) and the Michelson Doppler Imager (MDI).

As illustrated in Figure 3.1, EIT images show four temperature bands by capturing the Sun at four different wavelengths. The EIT 171 image in Figure 3.1.a is taken at 171 Angstrom corresponding to 1,000,000 degrees Kelvin. Images in Figure 3.1.b and Figure 3.1.c correspond to wavelengths of 195 and 284 Angstrom and temperatures 1,500,000 and 2,000,000 degree Kelvin, respectively. In Figure 3.1.d the image is taken at 304 Angstrom, corresponding to 60,000 to 80,000 degrees Kelvin.

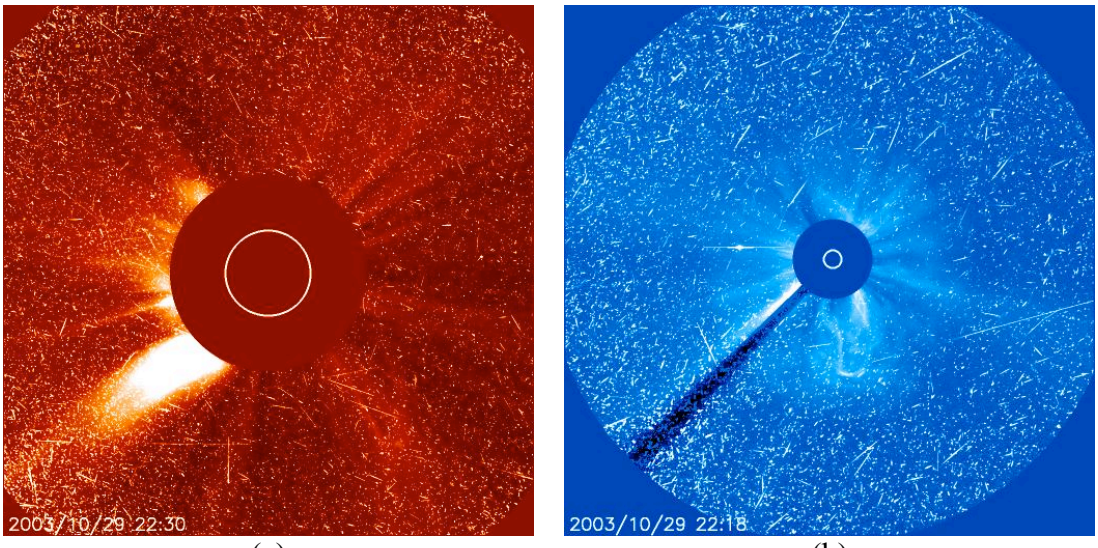
LASCO images are taken while blocking the light coming directly from the Sun which provides the ability to capture images of the solar corona. CMEs can be studied using the two LASCO coronagraphs: C2 and C3. Example of LASCO images are given in Figure 3.2 showing a fast halo CME which is ejected with linear speed of about 2029 km/s and hit the coronagraphs of SOHO. The C3 coronagraph has a large field of view providing images that cover a diameter of 45 million kilometres, while the C2 coronagraph provides images that cover the inner corona within 8.4 million kilometres of the Sun. This can be seen clearly by comparing the size of the solar disk shown in both images of Figure 3.2.



(a) (b)

(c) (d)

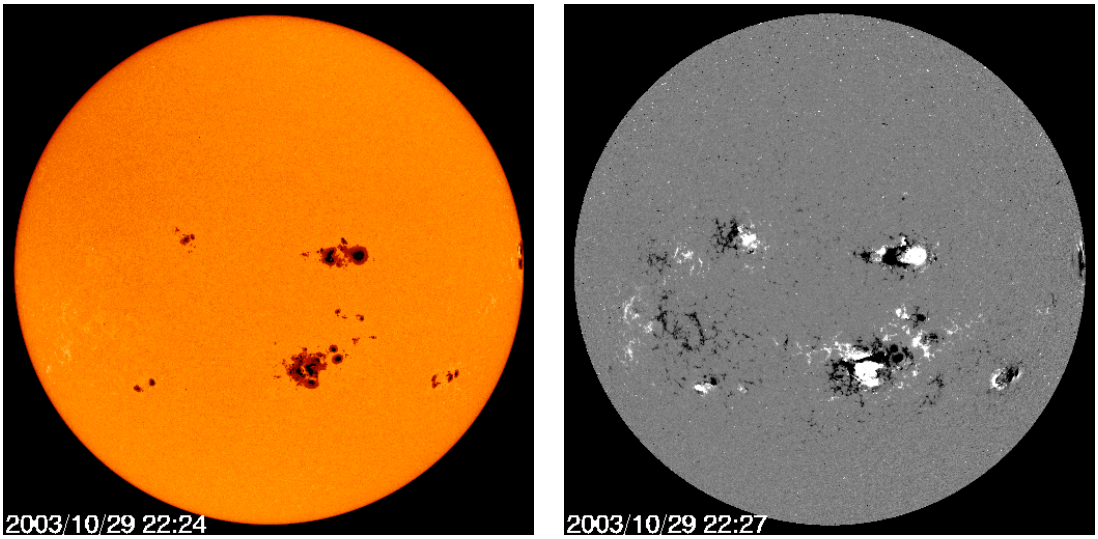
Figure 3.1 SOHO/EIT solar images were taken on 29/10/2003 (a) EIT 171 at 15:23UT (b) EIT 195 at 22:12UT (c) EIT 284 at 15:29UT (d) EIT 304 at 15:42UT.



(a) (b)

Figure 3.2 SOHO/LASCO solar images were taken on 29/10/2003 (a) LASCO C2 at 22:30UT and (b) LASCO C3 at 22:18UT.

Two types of MDI images are available: continuum and magnetogram images. A continuum image (Figure 3.3.a) shows white light intensity, which makes it easier to detect sunspots because they appear as dark spots due to their temperatures being lower than their surroundings. A magnetogram image (Figure 3.3.b) shows the horizontal component of the magnetic field and can be used to detect active regions as it depicts the line-of-sight component of the magnetic field of the solar photosphere.



(a) (b)

Figure 3.3 SOHO/MDI solar images were taken on 29/10/2003 (a) MDI Continuum at 22:24UT and (b) MDI Magnetogram at 22:27UT.

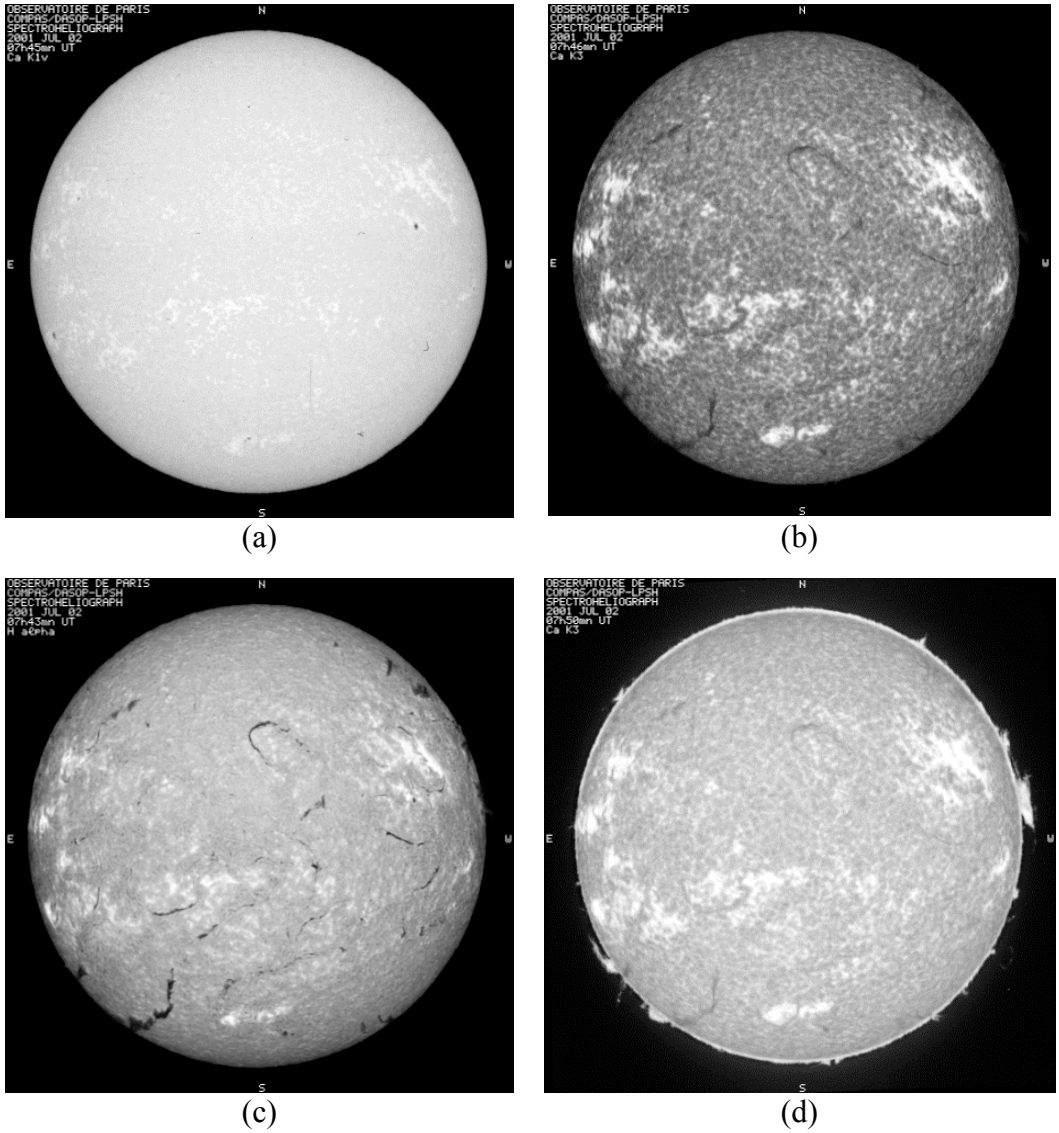
3.2.2 Ground-based telescopes

The Global High-Resolution Hydrogen-alpha Network[1] (GHN) provides access to many sources of solar images from different observatories around the world, such as the Mauna Loa Solar Observatory (MLSO) in Hawaii/USA, the Big Bear Solar Observatory (BBSO)[2] in California/USA, the Meudon Observatory in Meudon/France, the Pic du Midi Observatory in Tarbes/France, the Kanzelhohe Solar Observatory (KSO) in Treffen/Austria, the Catania Astrophysical Observatory (CAO) in Catania/Italy, the Huairou Solar Observing Station (HSOS) in Beijing/China and the

[1] http://www.bbso.njit.edu/Research/Halpha/, last access: 2009.
[2] http://www.bbso.njit.edu/pub/archive/, last access: 2008.

YunNan Astronomical Observatory (YNAO) in Kunming/China. For this work, solar data from Meudon Observatory[3] are used.

As shown in the examples of Figure 3.4, Meudon provides four main spectroheliograms: Ca II K1 images which are centred at 3933.2 Angstrom; Ca II K3v at 3933.7 Angstrom; Hydrogen-alpha (H-α) at 6562.8 Angstrom; and Ca II K3p prominences spectroheliogram centred at 3933.7 Angstrom.



(a) (b)

(c) (d)

Figure 3.4 Solar Images taken on 2/7/2001, provided by Meudon Observatory (a) Ca II K1v image at 7:45UT (b) Ca II K3 image at 7:46UT (c) H Alpha image at 7:43UT and (d) Ca II K3 prominences image at 7:50UT.

[3] http://bass2000.obspm.fr, last access: 2008.

The group of images shown in Figure 3.4 can be used for studying active regions and filaments/prominences. Filaments are defined as a mass of gas that can be seen in H alpha images as dark ribbons threaded over the solar disk (directly over magnetic-polarity inversion lines). If a filament is seen in emission against the dark sky (on the limb of the Sun) then it is called a prominence. An automated digital system was introduced in Mouradian (1998) to process Meudon's spectroheliograms and draw Synoptic maps of the solar activity as shown in Figure 3.5. The image of Figure 3.4.a is used to find the position and size of sunspot regions and that of Figure 3.4.b is used to draw contours of plages. Base-line of filaments is set based on the H-α image of Figure 3.4.c and the limb position of prominences is calculated from the image of Figure 3.4.d.

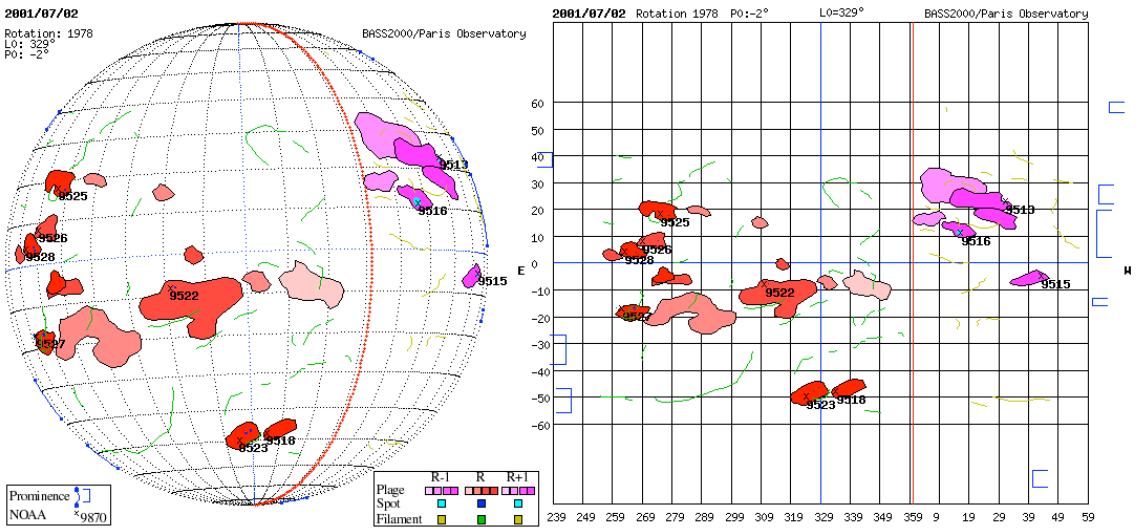


Figure 3.5 Synoptic maps of solar activity for the observations of Figure 3.4.

3.3 Data Catalogues

3.3.1 Sunspot Groups

Two catalogues for sunspot groups are used in the current work. The first catalogue is provided by the National Geophysical Data Centre[4] (NGDC). NGDC keeps a record of data from several observatories around the world and holds one of the most comprehensive publicly available databases for solar features and activities. The second

[4] ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/SUNSPOT_REGIONS/USAF_MWL/, last access: 2009.

catalogue is provided by the Space Weather Prediction Centre[5] (SWPC). SWPC monitors the solar and geophysical events with a real-time forecasting and it provides official space weather alerts and warnings.

The NGDC sunspot catalogue holds records of many solar observatories that have been tracking sunspot regions and supplying their date, time, location, physical properties, magnetic classification, sunspot area and the active region number (NOAA). A sample from this catalogue is shown in Figure 3.6.

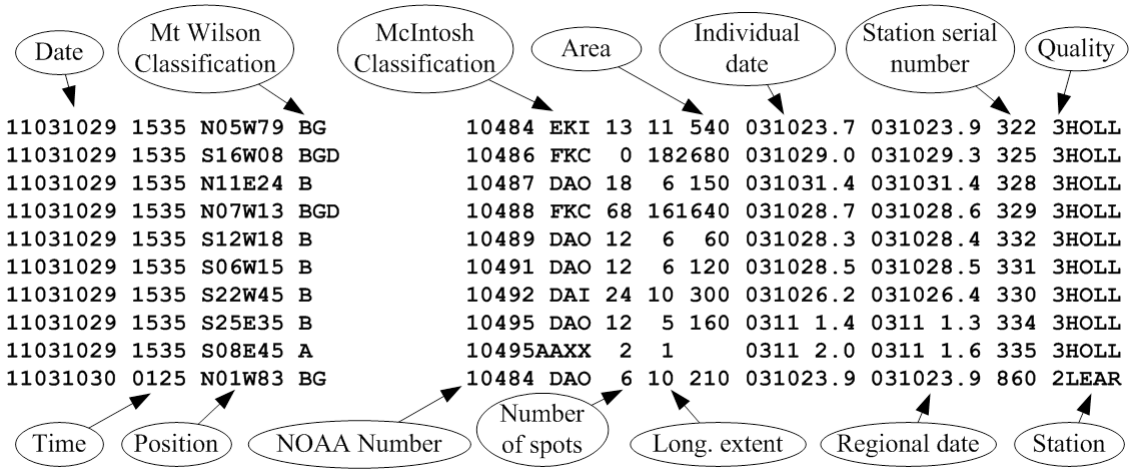


Figure 3.6 NGDC Sunspots Catalogue.

The SWPC sunspot catalogue holds records including dates, locations, area, extent, McIntosh class, active region numbers (NOAA), and the class of associated solar flare events. A sample from this catalogue is shown in Figure 3.7.

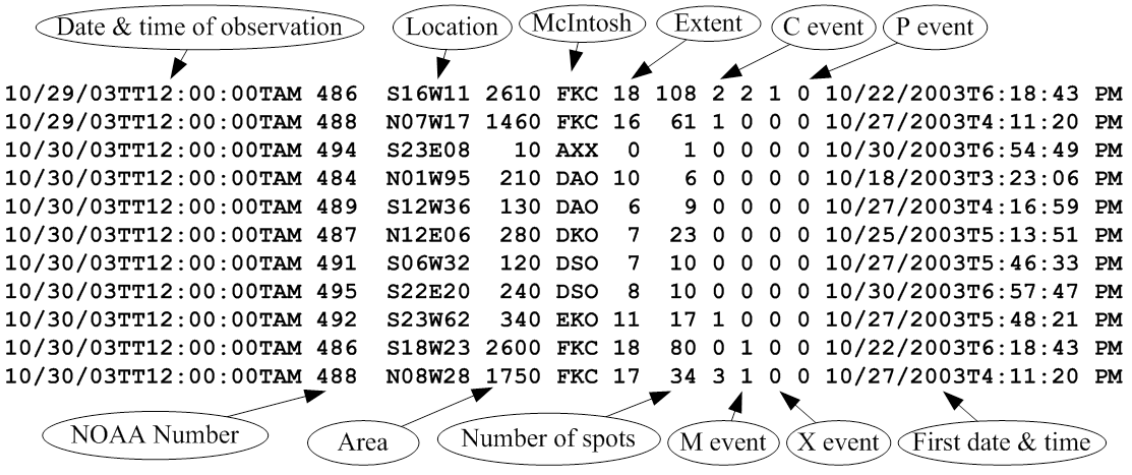


Figure 3.7 SWPC Sunspots Catalogue.

[5] http://www.swpc.noaa.gov/, last access: 2009.

In the sunspot catalogues, sunspot groups are classified using two methods: the Mount Wilson classification (based on magnetic configurations) and the McIntosh classification (based on visual properties).

The Mount Wilson (or Mt. Wilson) classification was first introduced by Hale et al. (1919) and consists of three major classes: Alpha (α: unipolar), Beta (β: bipolar), and Gamma (γ: complex). These classes are defined based on the distribution of magnetic polarities within spot groups (AFWAMAN15-1, 2003). In the cases where an inversion line separates umbrae of opposite polarity within the same penumbral area, an additional magnetic subclassification was introduced by Künzel (1960) which is the Delta (δ) configuration. According to rules set forth by the Mount Wilson Observatory in California, eight classes of Mt. Wilson are defined as described in Table 3-1.

Table 3-1 Mount Wilson magnetic classification system.

Mt. Wilson		Description
Class	Catalogue	
α	A	Unipolar: A sunspot group with one or more spots of the same polarity.
β	B	Bipolar: A sunspot group having both positive and negative magnetic polarities with a simple and distinct division between the polarities.
γ	G	Complex: A sunspot group in which the positive and negative polarities are so irregularly distributed as to prevent classification as a bipolar group.
δ	D	A qualifier to magnetic classes indicating that there are two or more umbrae of opposite polarity inside a single penumbra or penumbral area.
βγ	BG	A sunspot group that is bipolar but which is sufficiently complex that no single, continuous line can be drawn between spots of opposite polarities.
βδ	BD	A sunspot group of general beta magnetic classification but containing one (or more) delta spot(s).
βγδ	BGD	A sunspot group of beta-gamma magnetic classification but containing one (or more) delta spot(s).
γδ	GD	A sunspot group of gamma magnetic classification but containing one (or more) delta spot(s).

The McIntosh classification is a modified version of the Zurich classification system (Kiepenheuer, 1953) which has improved definitions and added indicators of size, stability and complexity. There are three components to the McIntosh classification system: the sunspot class; the penumbral class; and the sunspot distribution. According to McIntosh (1990), the general form of the classification is Zpc where, Z is the sunspot class which represents the modified Zurich class, p is the penumbral class which provides the type of the largest spot, and *c* is the sunspot distribution which represents the degree of compactness in the interior of the group. These components of the McIntosh classification system are described in Table 3-2.

Table 3-2 McIntosh physical classification system.

McIntosh		Description
Sunspot Class	A	Unipolar group with no penumbra. Length < 3° heliographic.
	B	Bipolar group with no penumbra. Length ≥ 3° heliographic.
	C	Bipolar group with penumbra on one end of the group, in most cases surrounding the largest of the leader umbrae.
	D	Bipolar group with penumbra on spots at both ends of the group. Length ≤ 10° heliographic.
	E	Bipolar group with penumbra on spots at both ends of the group. 10° < Length ≤ 15° heliographic.
	F	Bipolar group with penumbra on spots at both ends of the group. Length > 15° heliographic.
	H	Unipolar group with penumbra.
Penumbral Class	x	No penumbra.
	r	Rudimentary penumbra partially surrounds the largest spot.
	s	Small symmetric penumbra. The N-S (north-south) diameter across the penumbra ≤ 2.5° heliographic.
	a	Small asymmetric penumbra. The N-S diameter ≤ 2.5° heliographic.
	h	Large symmetric penumbra. The N-S diameter > 2.5° heliographic.
	k	Large asymmetric penumbra. The N-S diameter > 2.5° heliographic.
Spot Distribution	x	Undefined for a single spot or unipolar groups.
	o	Open. Few spots between leader and trailer.
	i	Intermediate. Numerous spots without a mature penumbra lie between the leading and trailing portions of the sunspot group.
	c	Compact. Many strong sunspots within the sunspot group, with at least one interior spot possessing mature penumbra.

There are some logical restrictions on combining the three components of the McIntosh classification system (AFWAMAN15-1, 2003). These restrictions limit the number of possible classifications in the system to 60 as shown in Table 3-3.

Table 3-3 Allowed types of groups in the McIntosh classification system; The Sunspot Distribution is shown for each allowed Sunspot-Penumbral Class pairs.

		Penumbral Class					
		x	r	s	a	h	k
Sunspot Class	A	x	-	-	-	-	-
	B	o, i	-	-	-	-	-
	C	-	o, i	o, i	o, i	o, i	o, i
	D	-	o, i	o, i, c	o, i, c	o, i, c	o, i, c
	E	-	o, i	o, i, c	o, i, c	o, i, c	o, i, c
	F	-	o, i	o, i, c	o, i, c	o, i, c	o, i, c
	H	-	x	x	x	x	x

3.3.2 Filaments/Prominences

Filament data from publicly available catalogues provided by the NGDC are used in this work. The NGDC filaments catalogue[6] holds records including dates, times, locations, physical properties, types, and active region numbers (NOAA) which have been supplied by many solar observatories around the world that have been tracking eruptive filaments/prominences. A sample from this catalogue is shown in Figure 3.8.

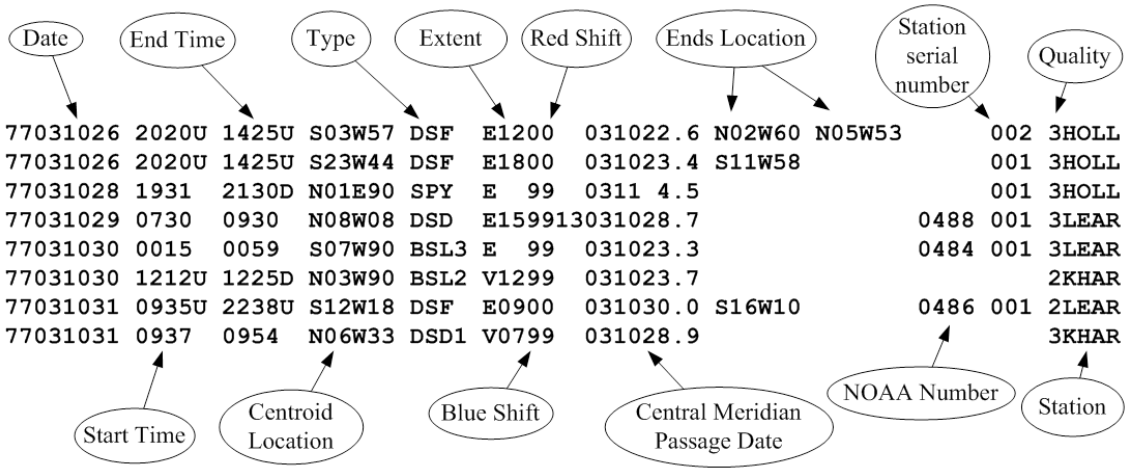


Figure 3.8 NGDC Filaments Catalogue.

[6] ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/SOLAR_FILAMENTS/, last access: 2008.

It is important to note that the start and end times of each filament in the catalogue are followed by a qualifier with three levels: D (after), E (before) and U (uncertain). In the catalogue, filaments are classified in 15 types as listed in Table 3-4.

Table 3-4 Filament types.

Type	Description
SSB	Solar Sector Boundary
MDP	Mound Prominence
CRN	Coronal Rain
CAP	CAP Prominence
LPS	Loops Prominence System
SPY	Spray
BSD	Bright Surge on Disk
APR	Active Prominence
DSD	Dark Surge on Disk
ADF	Active Dark Filament
ASR	Active Surge Region
AFS	Arch Filament System
BSL	Bright Surge on Limb
EPL	Eruptive Prominence on Limb
DSF	Disappearing filament

Two main types of filaments/prominences were first introduced by the Menzel-Evans scheme of classification (Menzel and Evans, 1953): (1) filaments originating in the coronal space and (2) filaments originating in the Chromosphere. Those originating from above in the coronal space consist of spot prominences (Loops and Funnels) and non-spot prominences (Coronal rain, Tree trunks, Trees, Hedgerows, Suspended clouds and Mounds). On the other hand, prominences originating from below in the Chromosphere include Surges and Puffs (spot) prominences and Spicules (non-spot) prominences. Detailed definitions of the filament types listed in Table 3-4 can be found in the glossaries provided by the Space Weather Prediction Centre (NOAA)[7] and the Space Environment Information System (SPENVIS)[8].

[7] http://www.swpc.noaa.gov/info/glossary.html, last access: 2009.
[8] http://www.spenvis.oma.be/spenvis/help/system/glossary.html, last access: 2009.

3.3.3 Solar Flares

As shown in Figure 3.9, the NGDC X-ray flares catalogue[9] provides information about the dates, starting and ending times for flare eruptions, locations, X-ray classifications, and the NOAA numbers for the active regions that are associated with the detected flares.

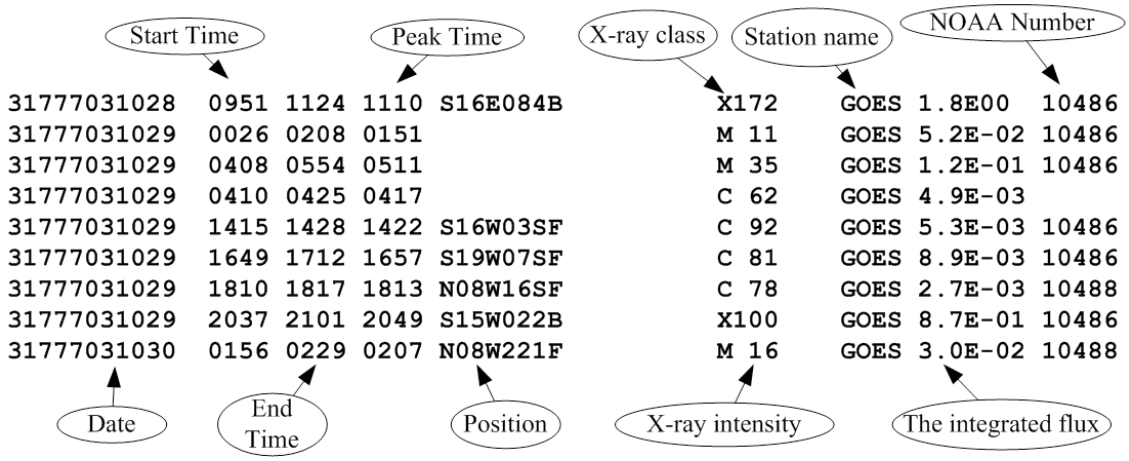


Figure 3.9 NGDC Flares Catalogue.

In this catalogue, flares are classified as A, B, C, M or X according to the peak x-ray emission (in watts per square meter, W/m²) in the 1-8 Angstroms wavelength band (Baker, 1970). In this classification, X class flares have a peak flux of order 10^{-4} W/m^2 and it is of order 10^{-5} W/m^2 for M class, 10^{-6} W/m^2 for C class, 10^{-7} W/m^2 for B class, and 10^{-8} W/m^2 for A class flares. So the X17.2 flare, included in Figure 3.9, would have an intensity of 17.2×10^{-4} W/m^2 and the C9.2 flare would have an intensity of 9.2×10^{-6} W/m^2.

3.3.4 CMEs

The CMEs data are obtained from the CME catalogue[10] that contains all CMEs manually identified since 1996 from the Large Angle and Spectrometric Coronagraph (LASCO) on board the Solar and Heliospheric Observatory (SOHO) mission (Gopalswamy et al., 2009a, Yashiro et al., 2004). This catalogue is generated and

[9] ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/SOLAR_FLARES/XRAY_FLARES/, last access: 2009.
[10] http://cdaw.gsfc.nasa.gov/CME_list/, last access: 2008.

maintained at the CDAW Data Centre by NASA and the Catholic University of America in cooperation with the Naval Research Laboratory. As indicated in Figure 3.10, this CME catalogue provides details of CME appearances, dates and times, position angles, angular widths, speeds and accelerations.

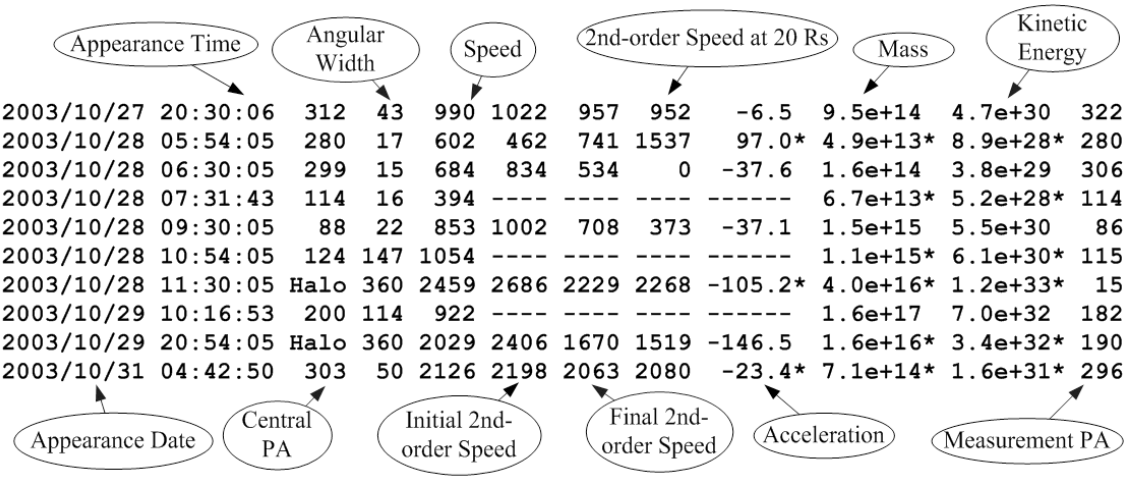
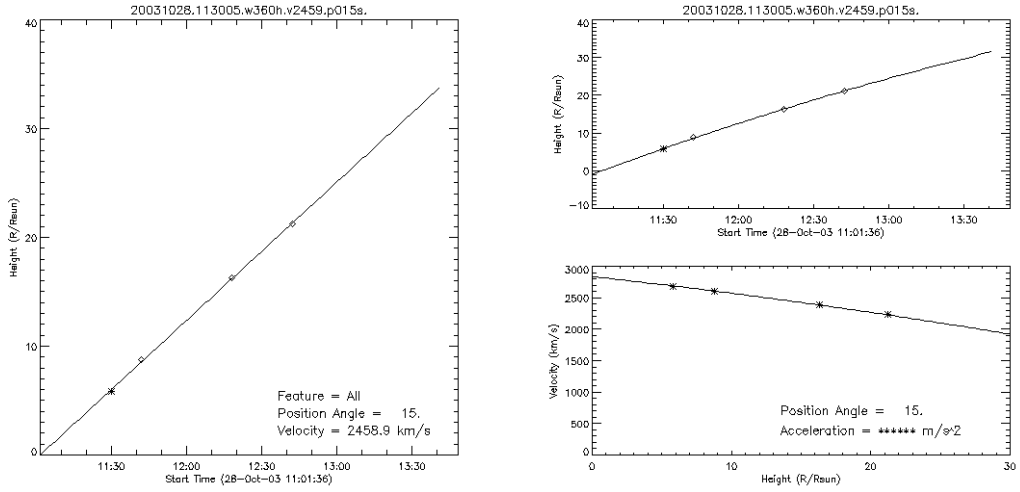


Figure 3.10 SOHO/LASCO CMEs Catalogue.

According to Yashiro et al. (2004) and Gopalswamy et al. (2009b), CME speeds are calculated by fitting a straight line to the height-time data as shown in Figure 3.11. Three speeds are calculated for each CME: a linear speed, quadratic speed at the time of the last possible height measurement, and quadratic speed at a height of 20 solar radii.



(a) (b)

Figure 3.11 Height-time measurement for a Halo CME which is recorded on 28 Oct 2003 at 11:30 (a) Linear fit (b) Second order fit.

3.4 Conclusions

Different sources of solar data were studied and are presented in this chapter. The conclusions drawn from this study highlight the following points:

- A total of 79304 sunspot groups are recorded in the NGDC sunspot catalogue in the period from 1996 to 2006. It is found that 18.7% of these sunspots are recorded without McIntosh classifications and 37% of them are reported without area.
- The SWPC sunspot catalogue provides one sunspot record a day per each active region. However, it is believed that one record a day is not enough to describe the variable physical properties and magnetic configurations within an active region along the day.
- A large number of filaments are missing from the NGDC filaments catalogue. This has been deduced by comparing the data in the filaments catalogue with the Synoptic maps produced by the Meudon Observatory. From the total number of reported filaments in the catalogue, for the years from 1996 to 2006 as listed in Table 3-5, it is clear that there are many data discrepancies including missing and repeated features.

Table 3-5 Total number of filament records per year as reported in the NGDC filaments catalogue.

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Number of Filaments	1989	2506	1320	446	593	479	412	373	165	132	102

- Out of 8765 filaments recorded in the NGDC filament catalogue in the period from 1996 to 2006, 68.3% are reported with the centroid location only (without the ends location) while 6.1% only are reported with the centroid and both ends location. On the other

hand, 17.5% of the filaments are reported without extent and 47.7% are reported without NOAA number.

- 19177 flares are reported in the NGDC flare catalogue in the period from 1996 to 2004. It is found that 50.2% out of these flares are recorded without locations and 47% of them are reported without NOAA number.
- Out of 11657 CME records, reported in the SOHO/LASCO CME catalogue in the period from 1996 to 2006, 21.4% of them are classified as poor events and 10.4% of them are marked as very poor events.
- The data gaps in the catalogues discussed in this chapter may affect the outcome accuracy of the association studies that are described in Chapter 4.

CHAPTER FOUR

4 DEVELOPING COMPUTERISED TOOLS TO FIND THE ASSOCIATIONS AMONG SOLAR ACTIVITIES

4.1 Introduction

Because of the effects that solar activities could have on our lives, it is becoming very important to understand the association between solar features and activities. This chapter describes automated computer algorithms developed to process years of historical solar data for sunspot groups, solar flares, filaments/prominences, and CMEs. The techniques presented in this chapter are used to study most of solar cycle 23 (a 10 year period). This enables us to study the activities during solar maximum and minimum periods and to study the associations with other activities or features.

Because of the huge amount of solar data available, it is impossible to study manually the associations mentioned above, which creates a need for automated association and data processing algorithms and tools. It is shown in the literature review of Chapter 2 that the associations between solar features and activities can be studied by overlaying different types of solar images. As shown in the previous chapter, there are many types of solar images that can be used for this purpose which means a need for complex image processing techniques with a capability for processing large size images. Since the NGDC sunspot, flare, and filament catalogues and the SOHO/LASCO CME catalogue (described in the previous chapter) are created manually based on solar

activity measurements that are calculated from solar images, it was decided to work on the associations that can be extracted from these catalogues.

This chapter introduces many algorithms associating flares and sunspots, CMEs and flares, CMEs and filaments and CMEs and sunspot-associated flares. A computer platform is designed to implement these association algorithms as one computer tool. This tool will extract the associations among solar activities and hence help to improve the understanding of many initiation models reported in the literature. This tool is also used in the research presented in this thesis to extract association datasets that can be processed by machine learning algorithms for the purpose of space weather forecasting as explained in Chapter 5.

This chapter is organized as follows: Section 4.2 introduces the CME-flare association algorithm. The CME-filament association algorithm is described in Section 4.3. The development of a generalized association algorithm to investigate the associations among CMEs, flares, and sunspot groups is described in Section 4.4. All the association algorithms introduced in this research were tested and the association results are presented in the sections 4.2 to 4.4. Conclusions on the association findings and comparisons with previous work are discussed in Section 4.5.

4.2 CME-Flare Associations

In this section, a C++ platform, created to automatically associate CMEs in the SOHO/LASCO CME catalogue with flares in the NGDC X-ray flare catalogue, is introduced. The association is determined on the basis of timing information, where the date and time for every CME is compared with date and time for every flare (Qahwaji et al., 2008c).

The algorithm starts by parsing the CME and flare catalogues. The association is based mainly on the flare-CME time line shown in Figure 4.1. "*A*" labels an associated flare, "*PA*" labels a possibly associated flare and "*NA*" labels a not-associated flare.

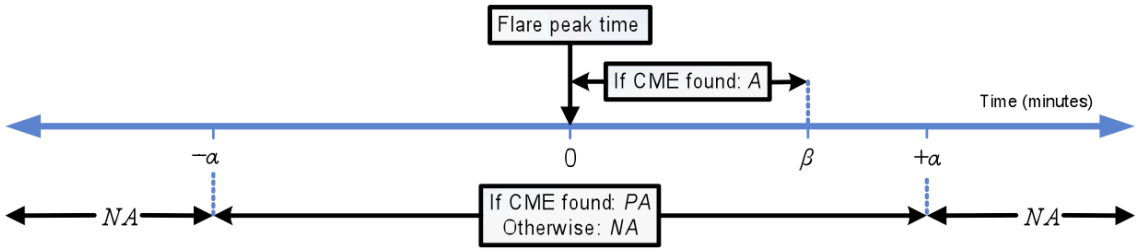


Figure 4.1 CME-flare time-based associations.

In Figure 4.1 two criteria are used for comparison:

- If there is no CME recorded "α" minutes before or after a flare reach its peak time, then this flare is marked as a *NA* otherwise it is marked as a *PA*.
- If there is a CME recorded "β" minutes after a *PA* flare reaches its peak then this flare is marked as an *A* flare.

The CME-flare association algorithm is used in Qahwaji et al. (2008c) to find the correlation between the data corresponding to 19164 solar flares and 9297 CMEs, which occurred during the period from 1 January 1996 to 31 December 2004.

Excluding 19 flaring events (2 M-class, 12 C-class, and 5 B-class flares), all the reported flares in the period of the study reached their peak intensities within less than 150 minutes. Hence, the value of α was made equal to 150 minutes in all the experiments to determine the *NA* and *PA* flares. It is easier to determine if a CME is not associated with any flares rather than determine the level of association between every CME with flares based on timing information. To explore the different levels of associations, the association algorithm was applied with different values of β processing X-, M-, C-, and B-class flares, as shown in Table 4-1.

As shown in Table 4-1, more CMEs are associated with flares as the value of β increases. The rate of increase in the number of associations is highest when β increases from 30 minutes to 60 minutes. The rates of increase are equal to 85%, 33% and 23% when β increases from 30 to 60, from 60 to 90 and from 90 to 120, respectively. Since

the increase in the association rate drops from 85% to 33% over a 60 minute difference, the value $\beta = 60$ was found to be most suitable for the analysis of associations.

Table 4-1 The numbers of *NA*, *PA* and *A* flares with different values of β for different classes of flares.

Flares	X	M	C	B	Total
NA	15	389	5554	3355	9313
PA (α=150)	89	926	6770	2066	9851
Total	104	1315	12324	5421	19164
A (β=30)	57	318	1181	246	1802
A (β=60)	71	510	2229	526	3336
A (β=90)	77	592	3016	764	4449
A (β=120)	78	654	3757	1018	5507

The CME-Flare algorithm was tested with $\alpha = 120$ and $\beta = 60$ and the associations dataset is created as shown in the sample of Table 4-2. The dataset provides flares' properties (date, time, class, and location) and properties of their corresponding CMEs (time, CPA, angular width, and speed).

By applying the association algorithm, with $\alpha = 150$ minutes and $\beta = 60$ minutes, to the data reported in the period between the years 1996 to 2004 a dataset of 581 associated flares was created. It is worth mentioning here that *PA* flares, those which have CMEs recorded before the flare peak time, were found and excluded from the association datasets so these datasets are more suitable to be processed by machine learning algorithms as explained in the next chapter.

4.3 CME-Filament Associations

Another C++ computer platform was created to automatically associate CMEs with eruptive filaments/prominences in the SOHO/LASCO and NGDC catalogues. The algorithm starts by parsing the CME and filament catalogues. Then a filament is labelled either "*A*", for associated, "*PA*" for possibly-associated filament, or "*NA*" for not-associated.

Table 4-2 Properties of flares and their associated CMEs.

Flare				CME			
Date	Time [UT]	Class	Location	Time [UT]	CPA [deg]	AW [deg]	Speed [km/s]
30/10/2003	16:14-16:24	C 5.7					
30/10/2003	18:30-18:43	C 5.8					
30/10/2003	19:18-19:27	C 5.6					
31/10/2003	01:50-01:56	C 5.5	N08W25				
31/10/2003	02:39-02:50	C 5.0	N08W33				
31/10/2003	04:26-04:37	M 2.0		04:42	303	50	2126
31/10/2003	06:08-06:28	M 1.1	N08W28	04:42	303	50	2126
31/10/2003	12:24-12:35	C 8.5					
31/10/2003	16:44-17:48	C 5.3		17:30	240	34	309
31/10/2003	20:14-20:49	C 5.1	N08W44				
31/10/2003	20:50-21:41	C 9.5	N08W44				
31/10/2003	23:50-00:37	C 4.4					
01/11/2003	04:17-04:26	C 3.7	S17W40				
01/11/2003	04:36-04:44	C 2.8					
01/11/2003	04:50-05:00	C 5.6	N09W47				
01/11/2003	05:26-05:34	C 3.6	S14W38				
01/11/2003	08:14-08:30	C 4.4	S13W41				
01/11/2003	08:39-09:06	M 1.3					
01/11/2003	11:31-12:04	C 9.7		12:30	263	68	246
01/11/2003	15:57-16:05	C 3.8	S16W45	14:54	274	55	334
01/11/2003	16:53-17:01	C 4.4	N08W49	14:54	274	55	334
01/11/2003	17:42-18:08	M 1.1	N09W50				
01/11/2003	19:12-19:18	C 3.5	N06W55				
01/11/2003	22:26-22:49	M 3.2	S12W60	21:30	318	143	413
01/11/2003	22:26-22:49	M 3.2	S12W60	23:06	254	93	899
02/11/2003	02:37-02:48	C 4.0	S14W52				
02/11/2003	06:59-08:12	M 1.0	S17W55				
02/11/2003	12:30-13:12	M 1.8		11:30	224	33	826
02/11/2003	17:03-17:39	X 8.3	S14W56	17:30	Halo	360	2598
03/11/2003	01:09-01:45	X 2.7	N10W83	01:59	304	65	827
03/11/2003	09:43-10:19	X 3.9	N08W77	10:06	293	103	1420
03/11/2003	15:26-15:43	M 3.9	S15W79				
03/11/2003	19:51-19:57	C 4.4					
03/11/2003	20:31-20:41	C 5.4					
03/11/2003	22:28-22:40	C 3.1					
04/11/2003	04:04-04:19	C 5.0					
04/11/2003	05:43-06:07	M 2.6					
04/11/2003	09:40-09:50	C 2.8					
04/11/2003	10:11-10:33	M 3.0					
04/11/2003	11:15-11:25	C 5.7					
04/11/2003	13:43-14:01	M 1.1		12:54	263	72	605
04/11/2003	19:29-20:06	X28.0	S19W83	19:31	197	52	327
04/11/2003	19:29-20:06	X28.0	S19W83	19:54	Halo	360	2657
05/11/2003	02:37-02:45	M 1.6	S19W89				

The associations between CMEs and filaments were first analysed for ten years of data (group 1: 1996-2006) and then the association algorithm was improved and applied to only six years of data (group 2: 1996-2001). This is explained in the following two subsections.

4.3.1 Group 1: Associations for Ten Years of Data (1996-2006)

The associations are determined as explained below:

- If a CME is recorded "α" minutes before or after the time a filament disappears, then this filament is labelled *PA*. Otherwise, it is labelled *NA* as depicted in Figure 4.2.

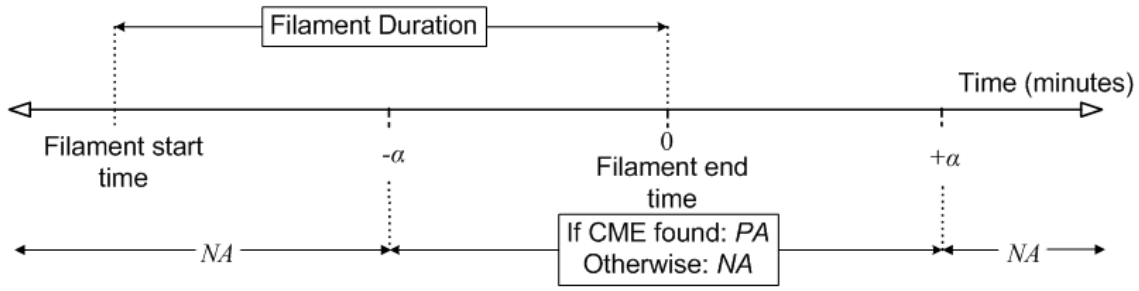


Figure 4.2 Time-based CME filament association (group 1).

- If the CPA for the recorded CME is located within ±30º of the centroid of the corresponding *PA* filament, as shown in Figure 4.3, then this filament is labelled *A* (Jing, 2005).

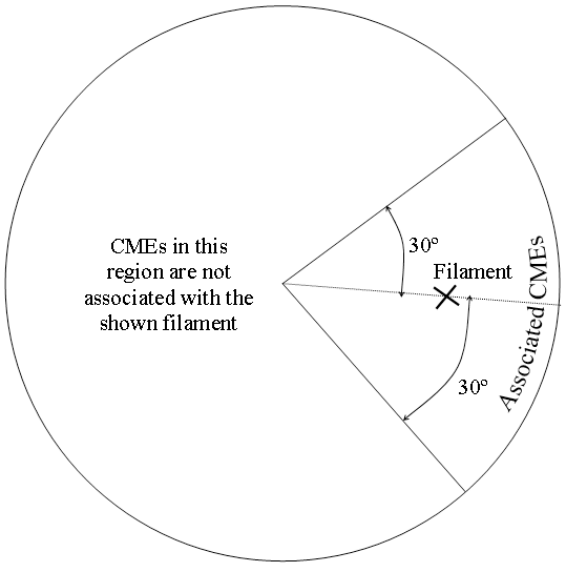
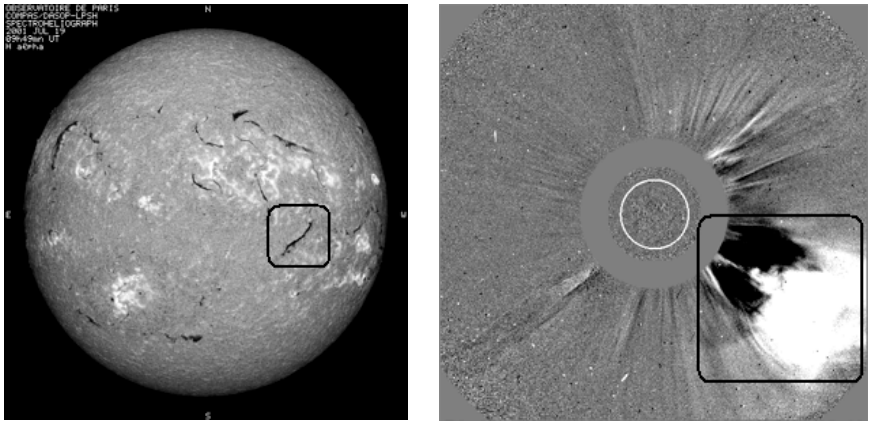


Figure 4.3 Location-based CME filament association.

For example, using these criteria we associate the filament of Figure 4.4.a and the CME of Figure 4.4.b as follows:



(a) (b)

Figure 4.4 (a) H alpha image taken on 19 Jul 2001 showing a filament with its centroid located in S20W59. (b) Time-differenced LASCO C2 image taken on 19 Jul 2001 showing a Partial Halo CME with central position angle of 275º.

- This filament, reported on 19 Jul 2001, started at 9:40 and disappeared at 10:15. The CME is first recorded on the same day at 10:30 (15 minutes after the disappearance of the filament).
- The filament is centred at S20W59, which produces an angle of 251º when converted to polar coordinations. The CME has a central position angle of 275º which falls within the filament association region (see Figure 4.5). Hence, this filament is labeled as associated (*A*) filament.

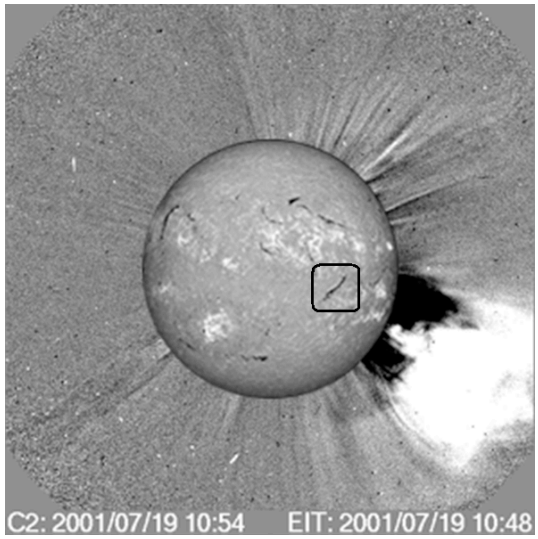


Figure 4.5 CME-filament association example.

For group 1, all the CME and filament data for solar cycle 23 in the period from January 1996 until the end of December 2006 were processed, resulting in the analysis of data relating to 11657 CMEs and 8765 eruptive filaments/prominences (Al-Omari et al., 2008, Qahwaji et al., 2008a).

To explore the different levels of association, the association algorithm was applied with different values of α, as shown in Table 4-3.

Table 4-3 Numbers of associations for different values of α.

α (minutes)	120	180	240	300
Associated filaments	928	1168	1396	1613
Associated CMEs	651	841	1009	1147
Distinct associations	70.2%	72.0%	72.3%	71.1%

As shown in Table 4-3, different values of α produce different levels of associations. The highest distinct associations percentage (without repeated associations) was obtained for α = 240 minutes, although this was not very different to the result for α = 180 minutes. Applying the association algorithm, with α = 240 minutes, an association dataset consisting of 2776 filaments with 1396 *A* filaments and 1380 *NA* filaments was created.

As concluded in the previous chapter, a large number of filaments are missing from the NGDC filament catalogue, especially for the years 2002-2006. This has been verified by comparing the data in the filament catalogue with Synoptic maps from Meudon Observatory. So, to better represent the associations between CMEs and filaments, the analysis was repeated and improved by considering six years of data only, the years 1996-2001, which form group 2.

4.3.2 Group 2: Associations for Six Years of Data (1996-2001)

As discussed in the previous subsection, the data of this group is selected as a better representation of the associations between CMEs and filaments. For this group, the associations are determined as discussed in the following four steps:

1. *Time-based associations*. The date and time of every CME are compared with the date and time of every filament (Al-Omari et al., 2008, Qahwaji et al., 2008a). The association labelling starts with the time-based associations. The CME event time is taken directly from the SOHO/LASCO CME catalogue. However, as most of the filament start and end times are reported in the NGDC filaments catalogue as uncertain, the average of the filament start and end times is taken to be the filament event time (Moon et al., 2002). As indicated in Figure 4.6, the width of the time association window is defined to be 2α minutes. If a CME is not recorded in the interval from α minutes before to α minutes after a filament event time, the filament is labelled *NA*; otherwise, it is labelled *PA* and recorded together with the relevant CMEs. To make the data sampling as homogeneous as possible, the value of α was the same in all association experiments and chosen to be 60 minutes, following Moon et al. (2002).

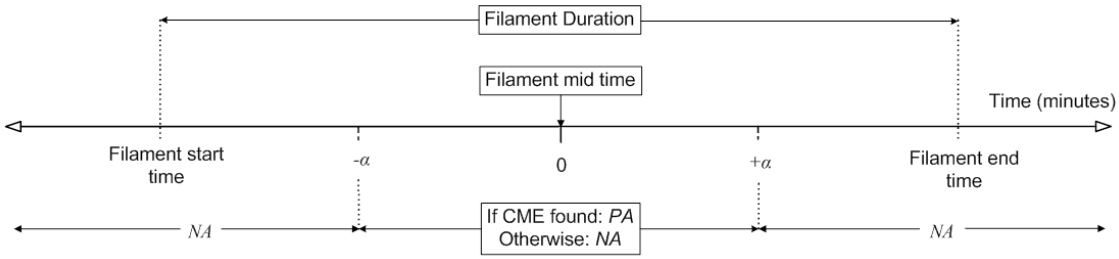


Figure 4.6 Time-based CME filament association.

2. *Location-based associations*. The Central Position Angle (CPA) of every CME is compared with the polar position of the centroid of every filament (Al-Omari et al., 2008, Qahwaji et al., 2008a). In this step, the algorithm analyses the *PA* filaments, identified by step 1, and the corresponding CME candidates. The algorithm defines an association sector on the solar disk within ±30º of the centroid of each *PA* filament as shown in Figure 4.3. If any of the CME candidates of a *PA* filament has a CPA lying within a filament's association sector, the filament is given the label *A* and recorded together with its associated

CME. In the cases where the candidates are halo CMEs, the Measurement Position Angle (MPA) is used instead because there is no CPA for a halo CME. According to Yashiro et al. (2004) and Gopalswamy et al. (2009b), MPA is defined for the fastest moving part of the CME's leading edge as the position angle at which the height-time information are measured. Except for CMEs that have a non-radial movement, the CPA and MPA are equal (Gopalswamy et al., 2009b). So, MPA can be used as an indicator of the CPA.

3. *Refining associations based on a CME's speed and acceleration*. According to Sheeley et al. (1999), CMEs can be classified into two classes: gradual and impulsive. The gradual CMEs are accelerating, with speeds ranging between 400 and 600 km/s and are associated with eruptive activities. The impulsive CMEs are decelerating, from speeds faster than 750 km/s and are initiated by solar flares. It is reported in Moon et al. (2002) that the median acceleration and speed for CMEs associated with significant flares (M and X classes) equal -8m/s² and 636m/s, respectively. Such CMEs can be assumed to be impulsive CMEs. By examining the CME acceleration and speed distributions for the filament-associated events in steps 1 and 2, it is found that these CMEs have zero median acceleration and a median speed of 417.5km/s as shown in Figure 4.7. As our algorithm associates CMEs with eruptive filaments/prominences, it is dealing with gradual CMEs (Sheeley et al., 1999). It was therefore decided to apply stricter association conditions that could lead to more accurate knowledge extraction with better machine learning performance (the work in Chapter 5). By making a simple comparison between the statistics of gradual and impulsive CMEs in the sample of data and those of Moon et al. (2002), it is clear that all CMEs that have accelerations less than -8m/s² and speeds greater than 636km/s are more likely to be associated with significant solar flares. Hence, the

associations were refined by ignoring any *A* filament with associated CME having acceleration less than -8m/s² or speed greater than 636km/s.

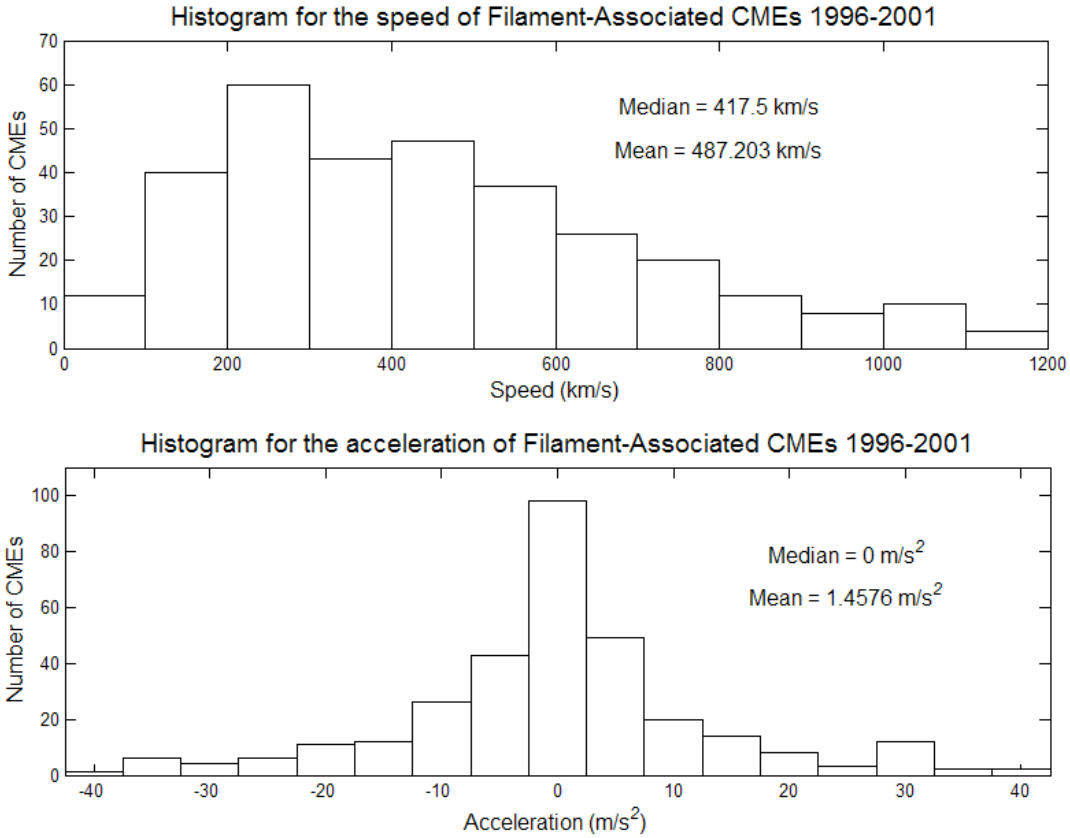


Figure 4.7 Distributions of speed and acceleration for filament-associated CMEs.

4. *Manual refinement*. By examining results from the association algorithm incorporating the previous three steps, it is apparent that the number of associated filaments can be greater than the number of associated CMEs which means that a single CME could have been associated with more than one filament. One should also consider the possibility of the data sets including single filaments that are each associated with more than one CME. These cases have been dealt with in the following way:

    - If a filament has more than one CME candidate then the algorithm will associate it with the CME closest in time and discard the rest.

- If the same CME is associated with many filaments then the case is investigated manually using H-alpha solar images that are obtained from Meudon Observatory and the BBSO. Such filaments are compared according to their distance from the limb, angular distance from the CME, duration, and extent. It is assumed that the associated filament is likely to be the one furthest from the centre of the solar disk, nearest to the CME, with the longest time duration or alternatively the greatest spatial extent.

Returning to the example given in Figure 4.5, it is found that the event time for the marked filament is calculated to be 9:57:30. The CME was first recorded on the same day at 10:30, about 32 minutes after the filament event time, which falls within the filament time association window. So, this filament is labeled *PA*. The *PA* filament is centred at S20W59 (an angle of 251º in polar coordinations) and the CME has a central position angle of 275º which falls within the filament association region. Hence, the filament is labeled as an *A* filament. According to step 3 of the association algorithm, this associated CME-filament pair should be discarded because the CME has a linear speed of 1668km/s and an acceleration of -11.6m/s.

By applying step 1 of the association algorithm, a total of 6101 out of the 7332 filaments were classified as *NA* filaments based on their timing information (Al-Omari et al., 2009a, Qahwaji et al., 2008b). A total of 1231 filaments were classified as *PA* filaments with 866 CME candidates out of the 5449 events recorded in the CME catalogues. The *PA* cases were compared on the basis of their locations and only 465 filaments were reclassified as *A* filaments, together with 330 CME events.

After applying the conditions relating to the speed and acceleration distributions of CMEs, in the third step of the algorithm, a total of 121 CME events have been

discarded so that only 209 out of the 5449 CMES (3.84%) are associated with a new set of 279 *A* filaments. Refining these association results manually, as described previously in step 4, resulted in the final classification from the association algorithm which is 209 *A* cases, 6101 *NA* cases, and 1022 *PA* cases (Al-Omari et al., 2009a).

The CME-filament association algorithm was tested with an association time-window of 4 hours width. The conditions related to location, speed, and acceleration were applied and a sample of the associated filaments dataset obtained is shown in Table 4-4. The dataset provides filament properties including date, time, duration, type, extent and location. At the same time it provides properties of the corresponding CMEs (time, CPA, angular width, and speed).

In the next chapter, these association datasets are used to define the learning rules that can be used later as part of an automated system for CME predictions based on filament observations. Datasets for the *NA* and *A* filaments are created representing their properties using a numerical format that is suitable for input to the machine learning algorithms. The *PA* filaments are excluded from the association datasets to make the learning performance as accurate as possible.

4.4 Associations among Sunspots, Flares and CMEs

A special algorithm is described in this section to find the associations between sunspots and solar flares (Qahwaji et al., 2007a). The algorithm described in Subsection 4.2, for finding CME/flare candidates is also used to search for CME candidates for the sunspot-associated flares (Qahwaji et al., 2007b).

As shown in the flow chart of Figure 4.8, the algorithm starts by parsing solar flares data and sunspot groups data from the NGDC catalogues. The flares data are read line by line and if an X, M, or C class flare is found, identified by its NOAA number, then the algorithm will try to identify its associated sunspot as follows:

Table 4-4 Properties of filaments and their associated CMEs.

Filament						CME			
Date	Time [UT]	Duration [min]	Type	Extent	Location	Time [UT]	CPA [deg]	AW [deg]	Speed [km/s]
05/01/2003	07:27-08:12	45	BSL	0	S30E90	07:54	120	127	293
07/01/2003	09:14-00:48	934	DSF	13	N03W31	17:54	286	95	300
13/01/2003	16:36-17:00	24	EPL	10	S18W84	16:54	228	77	275
20/01/2003	10:20-23:56	816	DSF	29	N15E45	15:30	44	100	96
20/01/2003	14:25-07:14	1009	DSF	21	N25E63	21:30	58	166	555
21/01/2003	18:48-15:06	1218	DSF	39	N14E39	05:06	42	29	426
27/02/2003	12:53-07:21	1108	DSF	20	N34E18	20:54	6	51	613
27/02/2003	12:53-07:21	1108	DSF	20	N34E18	22:30	55	13	0
21/03/2003	09:15-09:55	40	DSF	13	N12E59	10:54	54	66	481
01/04/2003	16:32-05:09	757	DSF	13	N36W23	22:30	310	63	177
04/04/2003	19:44-19:56	12	DSF	13	S03W43	21:19	291	89	487
11/04/2003	06:32-00:37	1085	DSF	10	S37W04	15:06	211	92	400
15/04/2003	17:17-11:38	1101	DSF	8	S22E61	04:06	86	23	363
18/04/2003	09:21-23:35	854	DSF	13	S25E21	14:50	118	8	549
18/04/2003	09:21-23:35	854	DSF	13	S25E21	14:50	161	98	139
27/04/2003	01:35-02:50	75	DSF	7	N12E57	03:26	50	82	547
07/05/2003	16:48-04:32	704	DSF	8	S31W17	23:06	191	17	484
07/05/2003	21:45-22:24	39	DSF	11	S34W14	23:06	191	17	484
15/05/2003	00:34-02:43	129	DSF	12	N18W88	02:54	303	26	342
11/07/2003	16:39-05:22	763	DSF	25	S05W26	00:30	256	15	397
12/07/2003	00:02-00:32	30	BSL	0	N01W72	00:30	256	15	397
15/08/2003	11:27-11:45	18	ADF	2	S09W11	13:31	231	38	153
25/08/2003	02:42-03:02	20	DSF	7	S09E38	03:26	106	61	575
25/08/2003	21:14-13:48	994	DSF	13	S07E52	04:50	120	32	322
25/08/2003	21:14-13:48	994	DSF	8	S11E44	04:50	120	32	322
07/09/2003	09:13-23:57	884	DSF	19	S20W23	16:30	200	51	324
07/09/2003	14:30-17:16	166	DSF	19	S38W18	14:30	227	72	376
07/09/2003	14:30-17:16	166	DSF	19	S38W18	16:30	200	51	324
07/09/2003	15:27-04:53	806	DSF	10	S25W26	21:54	223	91	227
10/09/2003	19:52-14:26	1114	DSF	13	N20W28	06:54	287	4	0
13/09/2003	16:20-05:41	801	DSF	9	N02W36	23:06	269	64	267
22/09/2003	10:28-10:42	14	DSD	3	N13E56	09:06	49	20	566
30/09/2003	08:44-09:05	21	DSD	12	N08W45	09:20	296	18	0
30/09/2003	10:30-10:39	9	ADF	4	S02W53	09:20	296	18	0
30/09/2003	20:26-20:34	8	EPL	0	S01E90	21:10	77	43	247
30/09/2003	20:50-21:05	15	EPL	0	S01E90	21:10	77	43	247
10/10/2003	11:06-11:20	14	EPL	0	S25W90	11:30	233	62	602
23/10/2003	07:19-22:38	919	DSF	23	N03W60	13:54	258	36	511
26/10/2003	00:50-01:45	55	EPL	10	S20W90	01:31	256	75	419
26/10/2003	02:07-03:24	77	DSF	14	S04W65	01:31	256	75	419
31/10/2003	09:35-22:38	783	DSF	9	S12W18	17:30	240	34	309
17/11/2003	22:50-14:24	934	DSF	17	S05E24	05:26	109	32	267
17/11/2003	23:23-14:20	897	DSF	9	S04E13	05:26	109	32	267
28/11/2003	02:54-03:45	51	DSF	11	S06W61	02:26	238	15	365

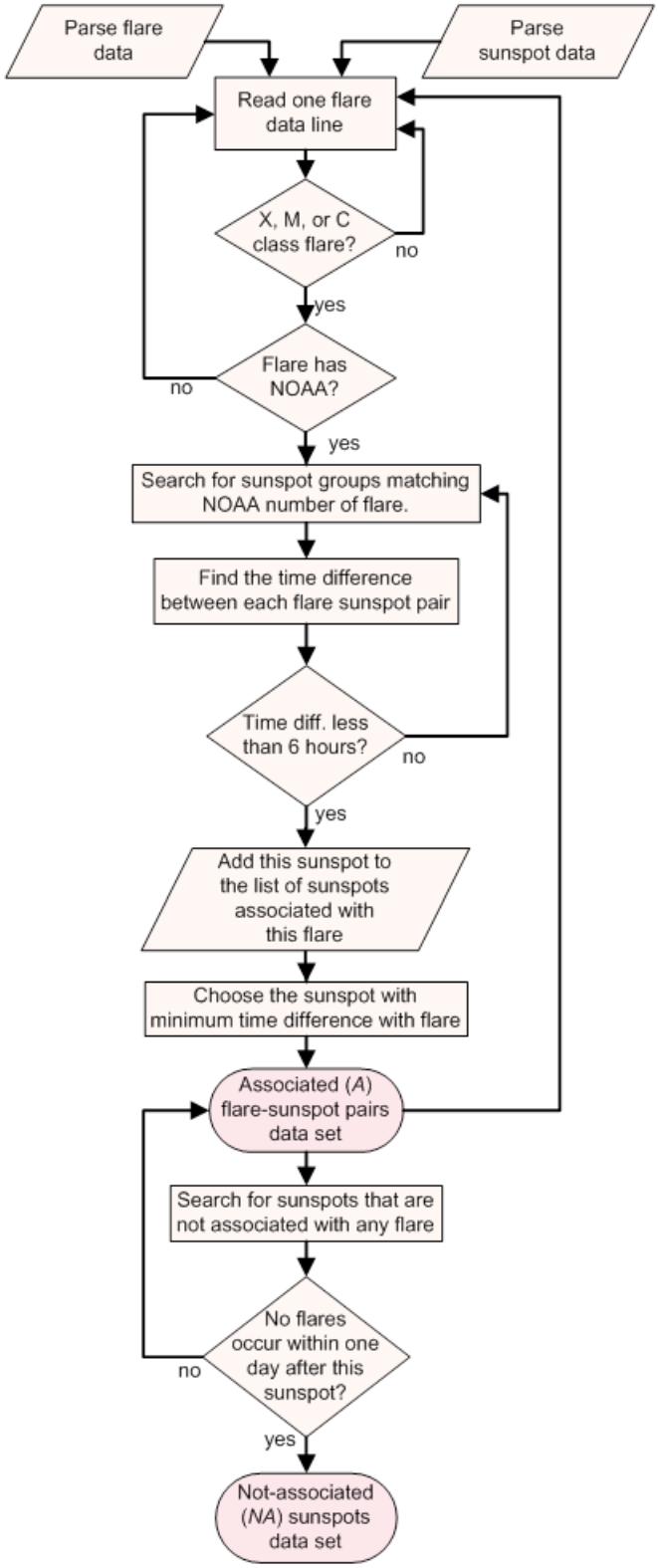


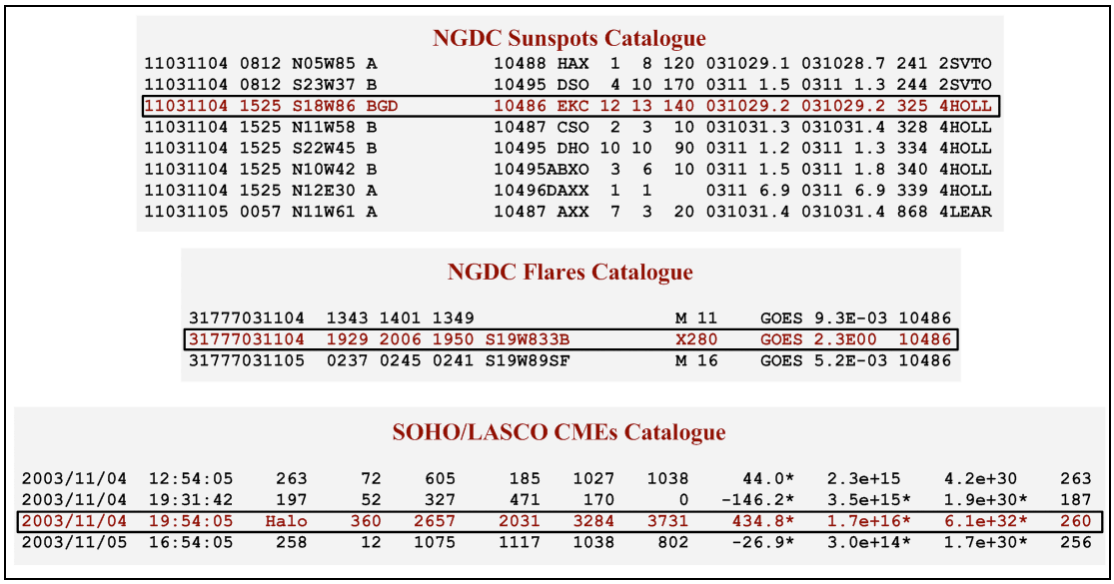
Figure 4.8 Flare-Sunspot Association Algorithm.

- Search the sunspot catalogue to find the sunspot groups with the matching NOAA number.
- Find the time difference between the flare eruption time and the observation time for sunspots.

- If the time difference is less than six hours then the flare and sunspot records are marked as being associated (*A*).
- If there is more than one classification report for the flare-associated sunspot group within six hours then the one with the minimum time difference is chosen.

The algorithm also collects the sunspot groups that are not associated with any flares. If no solar flares occur within one day after the classification of this sunspot group then it is marked as not-associated (*NA*).

After creating the sunspot-flare associated dataset, the algorithm described in Subsection 4.2 searches for CME candidates for the associated flares. In the association example shown in Figure 4.9, a sunspot is recorded on the 4th November 2003 at 15:25 within the active region 10486. At 19:29 an X28 flare started, which can be associated with the marked sunspot. Searching in the CME catalogue, it is found that a fast CME was recorded at 19:54.



NGDC Sunspots Catalogue

```
11031104 0812 N05W85 A          10488 HAX  1  8 120 031029.1 031028.7 241 2SVTO
11031104 0812 S23W37 B          10495 DSO  4 10 170 0311 1.5 0311 1.3 244 2SVTO
11031104 1525 S18W86 BGD        10486 EKC 12 13 140 031029.2 031029.2 325 4HOLL
11031104 1525 N11W58 B          10487 CSO  2  3  10 031031.3 031031.4 328 4HOLL
11031104 1525 S22W45 B          10495 DHO 10 10  90 0311 1.2 0311 1.3 334 4HOLL
11031104 1525 N10W42 B          10495ABXO  3  6  10 0311 1.5 0311 1.8 340 4HOLL
11031104 1525 N12E30 A          10496DAXX  1  1     0311 6.9 0311 6.9 339 4HOLL
11031105 0057 N11W61 A          10487 AXX  7  3  20 031031.4 031031.4 868 4LEAR
```

NGDC Flares Catalogue

```
31777031104  1343 1401 1349                  M 11    GOES 9.3E-03 10486
31777031104  1929 2006 1950 S19W833B         X280    GOES 2.3E00  10486
31777031105  0237 0245 0241 S19W89SF         M 16    GOES 5.2E-03 10486
```

SOHO/LASCO CMEs Catalogue

```
2003/11/04  12:54:05   263    72   605   185  1027  1038    44.0*  2.3e+15  4.2e+30  263
2003/11/04  19:31:42   197    52   327   471   170     0  -146.2*  3.5e+15* 1.9e+30* 187
2003/11/04  19:54:05  Halo   360  2657  2031  3284  3731   434.8*  1.7e+16* 6.1e+32* 260
2003/11/05  16:54:05   258    12  1075  1117  1038   802   -26.9*  3.0e+14* 1.7e+30* 256
```

Figure 4.9 Association Example, Nov, 4th 2003.

The algorithms were applied to sunspot, flare and CME data in the period from 1996 to 2006. Flare candidates were found within six hours after the sunspot observation time. Then, CME candidates were found within two hours before or after

the peak time of each associated flare. A sample of the association dataset obtained is given in Table 4-5 showing properties for the associated sunspots (date, time, Mount Wilson classification, McIntosh classification, and NOAA number), properties of the associated flares (time and class), and properties of the corresponding CMEs (time, CPA, angular width, and speed).

The associations described in this section can be used for the purpose of flare prediction, but in the current research it is used mainly, as described in the next chapter, for investigating the associations among sunspots, flares and CME for the purpose of CME predictions.

4.5 Discussions and Conclusions

The main aim of the work presented in this chapter was the creation of computer tools that can provide association datasets representing the relation between CMEs and flares and between CMEs and filaments. Because of the fixed data formats of the NGDC and SOHO/LASCO catalogues, they were found to be easy to access, analyse, and process in the current work. The association tool can process all the available data for 10 years of solar activity in less than 12 seconds on a 2 GHz processor computer. An exception occurs in processing the NGDC sunspots catalogue because it is found that the records in this catalogue are not sorted according to the solar cycle time, so more processing time is required to sort the sunspot records.

In Section 4.2, the association algorithm managed to associate 17.4% of the flares in the period from January 1996 to December 2004 with CME candidates. The association rates are found to be 68.3%, 38.8%, 18.1, and 9.7% for X, M, C, and B-class flares, respectively. It is clear that high association rates can be achieved if CME associations only with X and M class flares are considered. This seems appropriate for studying extreme space weather conditions because such significant flares can have hazardous impacts on human life on Earth.

Table 4-5 Properties of sunspot groups and their associated solar flares and/or CMEs.

Sunspot					Flare		CME			
Date	Time [UT]	MW	MI	NOAA	Time [UT]	Class	Time [UT]	CPA [deg]	AW [deg]	Speed [km/s]
26/10/03	02:37	BGD	FKC	10486	05:57-07:33	X 1.2	06:54	108	207	1371
26/10/03	15:09	BGD	EKC	10484	17:21-19:21	X 1.2	17:54	270	171	1537
27/10/03	00:22	BGD	DKC	10484	04:12-05:08	M 1.2	04:30	303	48	481
27/10/03	00:22	BGD	FKC	10486	01:33-01:44	C 6.2				
27/10/03	00:22	BGD	FKC	10486	06:13-06:28	C 9.0				
27/10/03	10:20	B	FKC	10486	12:27-12:52	M 6.7				
27/10/03	10:20	B	DKO	10488	14:02-14:21	C 5.1				
27/10/03	10:20	B	DKO	10488	14:53-15:07	C 5.7				
27/10/03	16:08	BGD	DKC	10484	19:48-20:16	C 9.0	20:30	312	43	990
27/10/03	16:08	BGD	FKC	10486	18:34-18:55	C 7.5	20:30	312	43	990
27/10/03	16:08	B	DKI	10488	21:46-22:05	M 1.9				
28/10/03	01:08	BGD	DKC	10484	05:07-05:14	C 7.7	05:54	280	17	602
28/10/03	01:08	BGD	FKC	10486	01:27-01:45	C 7.5				
28/10/03	07:49	B	FKC	10486	09:51-11:24	X17.2	10:54	124	147	1054
28/10/03	07:49	B	EKC	10488	08:35-08:44	C 8.7				
29/10/03	01:00	BGD	FKC	10486	04:08-05:54	M 3.5				
29/10/03	15:35	BGD	FKC	10486	16:49-17:12	C 8.1				
29/10/03	15:35	BGD	FKC	10486	20:37-21:01	X10.0	20:54	Halo	360	2029
29/10/03	15:35	BGD	FKC	10488	18:10-18:17	C 7.8				
30/10/03	01:25	BGD	FKC	10488	01:56-02:29	M 1.6				
30/10/03	07:50	B	FKC	10488	12:45-12:57	C 7.3				
30/10/03	07:50	B	EKO	10492	08:32-08:44	C 7.7				
31/10/03	00:58	BGD	FKC	10488	01:50-01:56	C 5.5				
31/10/03	00:58	BGD	FKC	10488	02:39-02:50	C 5.0				
31/10/03	00:58	BGD	FKC	10488	06:08-06:28	M 1.1				
31/10/03	15:38	BGD	FKC	10486	16:44-17:48	C 5.3	17:30	240	34	309
31/10/03	15:38	BGD	FKC	10488	20:14-20:49	C 5.1				
31/10/03	15:38	BGD	FKC	10488	20:50-21:41	C 9.5				
01/11/03	00:33	BGD	FKC	10486	04:17-04:26	C 3.7				
01/11/03	00:33	BGD	FKC	10486	05:26-05:34	C 3.6				
01/11/03	00:33	BGD	FKC	10488	04:50-05:00	C 5.6				
01/11/03	10:27	B	FKC	10486	15:57-16:05	C 3.8				
01/11/03	10:27	B	FKI	10488	11:31-12:04	C 9.7	12:30	263	68	246
01/11/03	16:15	BGD	FKC	10488	16:53-17:01	C 4.4				
01/11/03	16:15	BGD	FKC	10488	17:42-18:08	M 1.1				
01/11/03	16:15	BGD	FKC	10488	19:12-19:18	C 3.5				
02/11/03	02:21	BGD	FKC	10486	02:37-02:48	C 4.0				
02/11/03	02:21	BGD	FKC	10486	06:59-08:12	M 1.0				
02/11/03	15:45	BGD	EKC	10486	17:03-17:39	X 8.3	17:30	Halo	360	2598
03/11/03	00:45	BGD	FKC	10488	01:09-01:45	X 2.7	01:59	304	65	827
03/11/03	15:16	BGD	EKC	10486	15:26-15:43	M 3.9				
04/11/03	08:12	A	HKX	10486	09:40-09:50	C 2.8				
04/11/03	08:12	A	HKX	10486	11:15-11:25	C 5.7	12:54	263	72	605
04/11/03	08:12	A	HKX	10486	13:43-14:01	M 1.1				
04/11/03	08:12	A	HAX	10488	10:11-10:33	M 3.0				
04/11/03	15:25	BGD	EKC	10486	19:29-20:06	X28.0	19:54	Halo	360	2657

Associating CMEs with significant flares is supported by the findings of Yashiro et al. (2005), where it was found that all CMEs associated with X-class flares are detected by LASCO, while almost half the CMEs associated with C-class flares are invisible. In Yashiro et al. (2005) the authors used the term "invisible" when referring to the extremely faint CMEs that cannot be observed even if the coronagraphs were observing the solar corona with good cadence. They also concluded that the CME association rate increases with the increase of the X-Ray brightness for flares starting from 20% for C-class flares (between C3 and C9 levels) and reaching 100% for huge flares (above X3 level). In addition, they found that faster (median 1556 km/s) and wider (median 244°) CMEs are associated with X-class flares while slower (432 km/s) and narrower (68°) CMEs are associated with disk C-class flares.

In the associations for group 2 (Subsection 4.3.2), the time and location-based association algorithm (first two steps of the algorithm) associated 6.1% of the reported CMEs in the period 1996 to 2001 with filaments. This result is comparable with that obtained by Moon et al. (2002) who reported that 4% of the CMEs in the period 1996 to 2000 were associated with filaments on the basis of time and location using the same time-window width of 2 hours. The authors of Zhou et al. (2003) reported that more than 94% of the halo CMEs, in the period from 1997 to 2001, were associated with eruptive prominences/filaments but it is impossible to compare this result with the present ones because these authors did not include all available CMEs in the period. Instead they only selected 197 front-side halo CMEs. Here, it is important to mention that the final association dataset for group 2 (after applying the conditions related to the speed and acceleration) contains only 16 halo CMEs (7.7%), where the MPA is used to provide an indicator for CPA. The importance of halo CMEs for space weather comes from the fact that they are the most geo-effective CMEs and they are more likely to affect the Earth.

The location-based association condition between CMEs and filaments (constant association sector width of 60º) could be unreliable when associating filaments with the CMEs that have larger angular widths. For this reason, the algorithm is checked using a dynamic association sector such that the sector width is set to 60º for CMEs with angular width < 60º and it is set to the angular width of the CME under consideration for CMEs with larger angular width. By applying the association algorithm again, the same association results (as the final classifications mentioned previously for group 2) were obtained plus an extra 21 associated CME events with angular width > 60º. Because of the large angular widths of these extra CMEs, they were associated with many filaments. For example, a partial halo CME was recorded on 19 Oct 1996 at 17:17 with angular width of 170º and CPA of 159º. This CME was associated with 4 filament records having the centroids S08E47, S09E41, S28E90 and S19E55. After checking H-alpha images it was found that these filaments have approximately the same angular distance of about 50º from the CPA of the CME and therefore it is difficult to decide which filament is the relevant one. For the purpose of making the machine learning study in the next chapter more accurate, it is preferred to exclude the extra 21 cases because it is believed that having a small dataset of correctly associated CME-filament pairs is better than having a larger dataset that contains some incorrectly associated pairs.

Another issue that should be mentioned here is that the SOHO/LASCO CME catalogue does not include any information to distinguish between front side and backside CMEs. Therefore, it is possible for the proposed tool to associate a backside CME with a flare or a filament.

The final association datasets in this chapter, which were obtained by analysing data catalogues, are processed by methods described in the next chapter using different

machine learning algorithms to extract the relation between solar activities and to provide computerised learning rules for the purpose of CME predictions.

CHAPTER FIVE

5 AUTOMATED PREDICTION OF SOLAR ACTIVITIES AND FEATURES USING MACHINE LEARNING

5.1 Introduction

This chapter introduces a machine learning-based computer platform for analysing the associations between CMEs, flares and filaments data within the context of CME predictions. Theoretically, machine learning algorithms have the potential to extract knowledge from the associated solar features and activities and represent this knowledge in computerised learning rules that can be used within the context of space weather prediction. The feasibility of using them for studying the associations is investigated in this chapter from a practical perspective. The training of the learning algorithms is based on the association datasets created in the previous chapter by investigating the CME associations with flares on one hand, and the CME associations with filaments/prominences on the other.

This chapter is organized as follows: A short description of the machine learning algorithms used in the current work is provided in Section 5.2. Verification methods, validation techniques and performance indicators are described in Section 5.3. The system design for CMEs off-line predictions is expressed in two sections: Section 0 discusses the predictions based on the associations between CMEs and solar flares, while Section 5.5 investigates the CME predictions based on their associations with

filaments/prominences. Section 5.6 describes a solar flare prediction platform provides a general study for the prediction of CMEs based on their associations with sunspots-associated flares. A comparison between the performances achieved is provided in Section 5.7. Finally, Section 5.8 draws some conclusions from the work presented in this chapter.

5.2 Machine Learning

In this work, Cascade-Correlation Neural Networks (CCNNs), Support Vector Machines (SVMs), Radial Basis Function Networks (RBFNs) and the Adaptive Boosting (AdaBoost) algorithm are used and compared for the purpose of CME predictions. Many existing references, such as Qu et al. (2003), Qahwaji and Colak (2007) and Qahwaji et al. (2008b), provide detailed description of these learning algorithms. So, this section will provide only short descriptions of them.

5.2.1 Cascade-Correlation Neural Networks (CCNNs)

All CCNN experiments were carried out using the MATLAB neural network toolkit. The number of input nodes in a CCNN is determined by the number of input features, while the number of output nodes is determined by the number of different output classes. The learning of CCNN starts with no hidden nodes. The direct input-output connections are trained using the entire training set with the aid of the back propagation learning algorithm. Hidden nodes are then added gradually and every new node is connected to every input node and to every pre-existing hidden node. Training is carried out using the training vectors and after each pass the weights of the new hidden nodes are adjusted (Fahlmann and Lebiere, 1989).

5.2.2 Support Vector Machines (SVMs)

The "MySVM" software (Rüping, 2000) was used for the SVM experiments. The Anova-Kernel SVM has been used as it was found to outperform the NNs used for

the solar data processing in (Qahwaji and Colak, 2007). The Anova kernel is defined by the sum of exponential functions in the x and y directions,

$$k(x,y) = \left[\sum_i \exp\left(-\gamma\left(x_i - y_i\right)\right)\right]^d \qquad (5-1)$$

where the parameters d (degree) and γ (gamma) control the shape of the kernel. Optimisation of the SVM performance was done by adjusting d, γ and the classification threshold. The classification threshold is simply the decision value at which the data can be classified into two classes. So, SVM classifications above this threshold would be associated with class 1 (positive prediction) and the rest of the data would be associated with class 2 (negative prediction).

5.2.3 Radial Basis Function Networks (RBFNs)

RBFNs are powerful interpolation techniques that can be efficiently applied in multidimensional space. The RBFN approach to classification is based on curve fitting. Learning is achieved when a multi-dimensional surface is found that can provide optimum separation of multi-dimensional training data. In general, RBFNs can model continuous functions with reasonable accuracy. The radial basis functions are the set of functions provided by the hidden nodes that constitute an arbitrary "basis" for the input patterns (Qu et al., 2003). One of the major advantages of using RBFNs is that their training is usually simpler and shorter than the training of other NNs. However, greater computation and storage requirements for classification of inputs are usually required after the network is trained (Sutton and Barto, 1998).

5.2.4 Adaptive Boosting (AdaBoost) Algorithms

The AdaBoost algorithm, described in Freund and Schapire (1997), was used in Qahwaji et al. (2008b) for CME prediction. It constructs a "strong" classifier using a training data set and a linear combination of weak hypothesis. Hence, the constructed

AdaBoost classifier "boosts" the weak classifiers to provide a stronger one. Three boosting algorithms were compared: Real, Gentle and Modest AdaBoost. Real AdaBoost is the boosting algorithm reported in Schapire and Singer (1999), which is a generalisation of the basic AdaBoost algorithm introduced in Freund and Schapire (1996). Gentle AdaBoost, introduced in Friedman et al. (2000), is a more robust and stable version of the Real AdaBoost algorithm and performs slightly better than the latter on regular data and considerably better on noisy data (Friedman et al., 2000). Modest AdaBoost, described in Vezhnevets and Vezhnevets (2005), can provide better generalization capability and higher resistance to over fitting compared to the alternative forms of AdaBoost. In addition, Modest AdaBoost, in certain cases, can provide good performance in terms of test error. The three AdaBoost algorithms have been implemented using the AdaBoost MATLAB toolbox[11]. This toolbox was designed by the Graphics and Media Lab (GML) at the department of computer science at Moscow State University.

5.3 Verification and Validation Techniques

5.3.1 The Jack-Knife Technique

The Jack-knife technique is used to provide a correct statistical evaluation of the performance of a classifier when it is trained and tested on a relatively limited number of samples. The technique divides the total number of samples into two sets: a training set and a testing set. In practice, a random number generator is used to divide the samples into these two sets. For a finite number of samples, an error counting procedure can be used to estimate the performance of the learning algorithms (Fukunaga, 1990).

The cross validation technique is not used in the work of this chapter because there are many more negative instances (*NA* events) than positive instances (*A* events) in the solar data used (Chapter 4). In addition the samples of solar data were sorted

[11] http://research.graphicon.ru/machine-learning/gml-adaboost-matlab-toolbox.html, last access: 2008.

according to the solar cycle timing information which increases the chance that a given subsample may not contain any CME-associated filaments or flares as there are no significant solar activities during the solar minimum; consequently, this will reduce the classifier training performance.

5.3.2 Performance Indicators

The following performance indicators are used: True Positive Rate (*TPR*), False Positive Rate (*FPR*), True Negative Rate (*TNR*), False Negative Rate (*FNR*), Accuracy, Specificity, Sensitivity, and Heidke Skill Score (*HSS*). Since the system design aims to predict if an eruptive filament or a solar flare is going to initiate a CME (positive) or not initiate a CME (negative), these indicators are defined as follows:

$$TPR = \frac{TP}{\text{Total actual positives (number of } A \text{ cases)}} = \frac{TP}{TP + FN} \qquad (5-2)$$

where *TP* (True Positives) is the total number of cases for which the system correctly predicts that a flare or a filament produces a CME and *FN* (False Negatives) is the number of cases where the system predicts incorrectly that a flare or a filament does not produce a CME,

$$FPR = \frac{FP}{\text{Total actual negatives (number of } NA \text{ cases)}} = \frac{FP}{FP + TN} \qquad (5-3)$$

where *FP* (False Positives) is the total number of cases for which the system predicts incorrectly that a CME is produced and *TN* (True Negatives) is the number of cases where the system predicts correctly that a CME is not produced,

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (5-4)$$

where the summation *TP*+*FP*+*TN*+*FN* is the total number of associated and not-associated events used in the experiments.

Specificity is an indicator of a system's ability to correctly identify negatives. From Equation 4-3 and the definition of *TN*, Specificity = 1−*FPR* = *TNR*. Sensitivity,

on the other hand, is an indicator of a system's ability to correctly identify positives and can be defined as the ratio of the number of true positives to the sum of true positives and false negatives or in other words, Sensitivity = *TPR*.

The Heidke skill score is reported in Balch (2008) and Heidke (1926) and defined as

$$HSS = \frac{TP + TN - E}{TP + FP + TN + FP - E} \qquad (5-5)$$

where E is the number of correct predictions which would be made by chance and is calculated as

$$E = \frac{(TP + FP)(TP + FN) + (FP + TN)(FN + TN)}{TP + FP + TN + FN} \qquad (5-6)$$

HSS ranges from -1 (which means all incorrect prediction) to +1 (which means all correct prediction). If a prediction system has a zero *HSS* score, then the system performance is no better than that from random guessing (Balch, 2008).

When testing the suggested prediction techniques, described in the sections to follow, most of the performance indicators described above were calculated. The most common way to evaluate the performance of a forecasting system is the use of Receiver Operating Characteristic (ROC) curves, as explained in Fawcett (2006). An ROC curve plots the *FPR* on the x-axis and the corresponding *TPR* on the y-axis such that the diagonal line corresponds to random guessing (Fawcett, 2006). According to Fawcett (2006), the system with best performance is the one on the ROC curves which is furthest from the diagonal line in the upper-left direction. Mathematically, if we have different systems/configurations and each one is represented on an ROC curve by a point (*FPR*,*TPR*) then the system/configuration point with the maximum distance to the diagonal line, in the upper-left direction, has the best performance. The distance D_{ROC} from a point (*FPR*, *TPR*) to the diagonal line can be expressed as:

$$D_{ROC} = \frac{|FPR - TPR|}{\sqrt{2}} \qquad (5-7)$$

Based on the definition of these indicators, it is clear that the *TPR*, *FPR*, and *HSS* are the most important indicators to measure the performance of a space weather forecasting system. It is not enough to obtain high *TPR* values and low *FPR* values, but also the *HSS* is needed to be as high as possible to ensure that the system is not predicting by chance.

5.4 CME Predictions Based on CME-Flare Associations

5.4.1 Data Handling

In Chapter 4, 71 X-class flares and 510 M-class flares were associated with CMEs, while 15 X-class flares and 389 M-class flares were not associated. Because machine learning algorithms deal mainly with numbers, it was essential that appropriate numerical representations for *A* and *NA* flares were used. This can be seen in Figure 5.1 which depicts the proposed CMEs prediction system (based on flare associations).

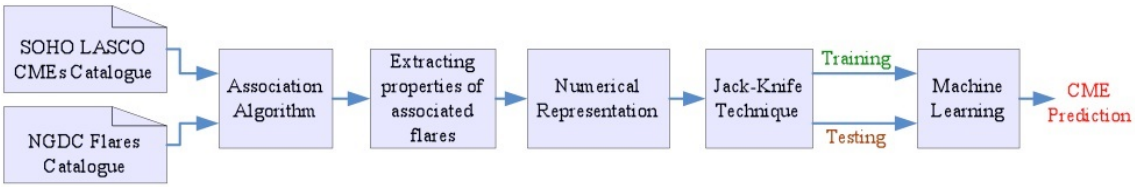


Figure 5.1 The hybrid prediction computer system.

Properties such as intensity, starting time, peak and ending time of the flares can be extracted from the NGDC flares catalogue. It was hoped to include additional properties for flare locations. Unfortunately a large number of the associated flares do not have locations included in the NGDC catalogues. Hence, it was decided to use the properties shown in Table 5-1.

Numerical representations of these properties are used to construct the input parameters for the training and testing stages of the machine learning system. The flare properties shown in Table 5-1 were calculated and normalised to be in the range

between 0.1 and 0.9. The single target function (output node) has a numerical value of 0.9 if a CME is likely to occur and 0.1 if not. To find which of the properties are the most significant for the prediction of CMEs using machine learning, extensive experiments were carried out in order to determine the significance of each feature for the proposed application.

Table 5-1 The features extracted for flares.

Flare property	Description
Intensity	The normalized numerical value for the intensity of the flare (Intensity × 1000).
Flare Duration	The normalized numerical value for the time difference (in minutes) between the ending and the starting time of the flare (Difference/120).
Decline Duration	The normalized numerical value for the time difference (in minutes) between the ending and the peak time of the flare (Difference/120).
Incline Duration	The normalized numerical value for the time difference (in minutes) between the peak and the starting time of the flare (Difference/120).

Both CCNNs and SVMs have proven to be very effective learning algorithms for similar applications (Qahwaji and Colak, 2007). So, their performances were compared in this section. The Jack-knife technique was used in all experiments with the use of 80% randomly selected samples for training and the remaining 20% for testing. The associated dataset created by the tools described in Chapter 4, consists of 985 flares with 581 *A* flares and 404 *NA* flares. Consequently, a total of 788 *A* and *NA* flares were used for training and 197 *A* and *NA* flares were used for testing. The prediction performances are evaluated using the ROC analysis technique with two indicators: *TPR* and *FPR*.

5.4.2 CCNN Experiments

In the CCNN experiments, the number of input parameters/nodes and the number of hidden nodes in each experiment were changed to find the best inputs and

their related topologies. The number of input parameters was varied from 1 to 3 and 20 CCNN configurations were created for each input feature by changing the number of hidden nodes from 1 to 20. Five experiments were carried out based on the Jack-knife technique for each CCNN configuration and the average *TPR* and *FPR* were recorded. At the end of these experiments, 60 CCNN topologies were compared by the ROC curve of Figure 5.2.

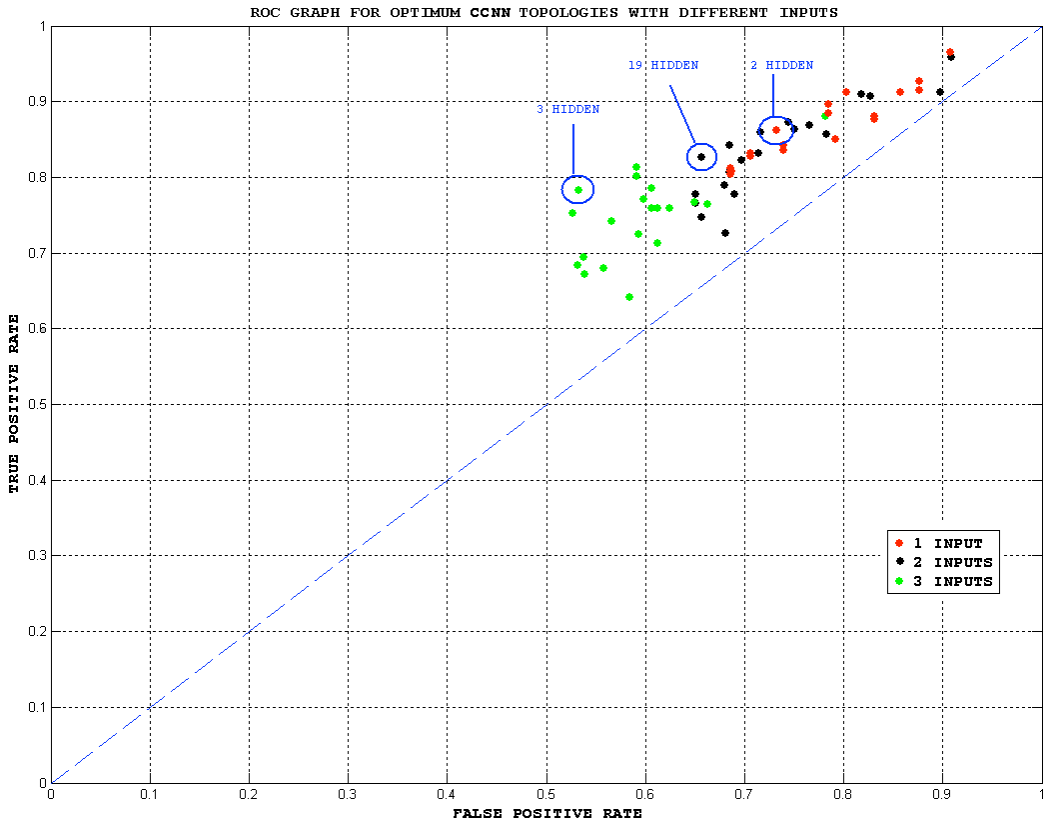


Figure 5.2 ROC graph showing the best CCNN topologies with different inputs.

The optimum topology for each number of input features was found as indicated in Figure 5.2 by determining the point with the maximum perpendicular distance from the diagonal line in the upper-left direction as explained in Section 5.3. Then for each of these optimum topologies, the optimum classification threshold values was found by changing the threshold values from 0 to 1 in steps of 0.01 for each input and their associated optimum topologies. For each threshold value, five experiments were carried out using the Jack-knife technique and the average *TPR* and *FPR* values were recorded and then another ROC curve was created as shown in Figure 5.3.

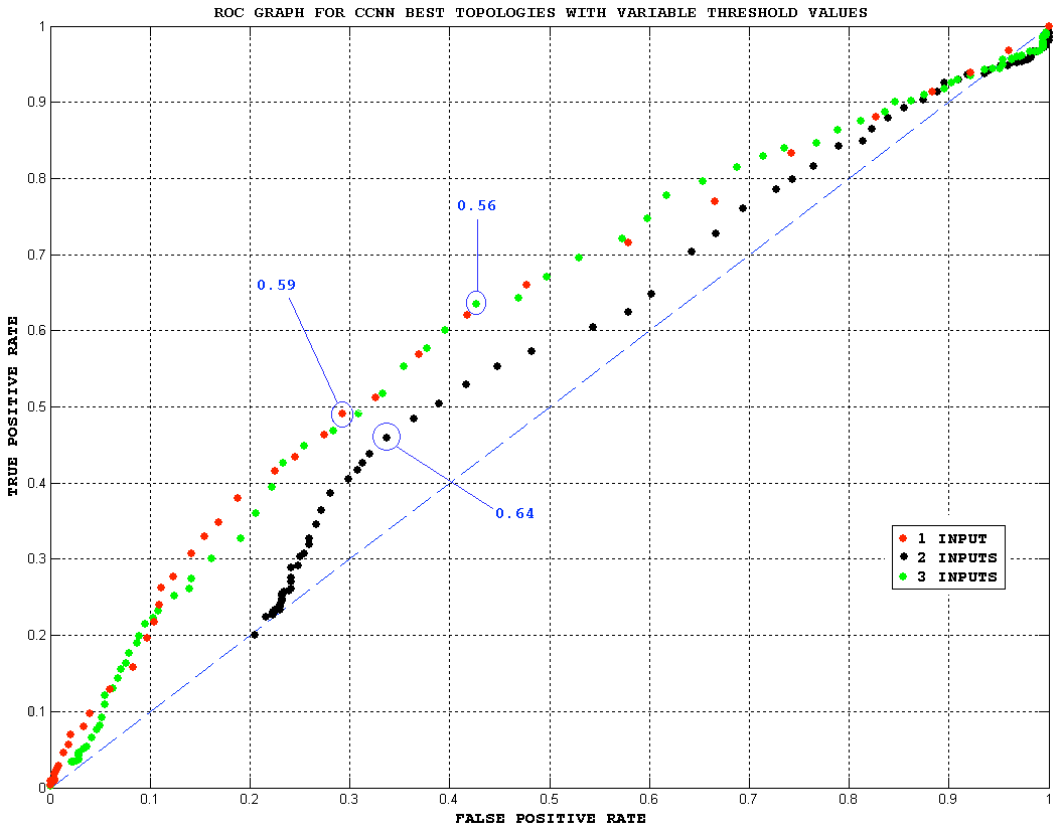


Figure 5.3 ROC graph showing the best CCNN topologies with different inputs and variable threshold values.

From Figure 5.3, it is found that a CCNN with 3 input nodes and 3 hidden nodes with a classification threshold of 0.56 gives the best results for CME prediction as it provides 0.63 *TPR* and 0.43 *FPR*.

5.4.3 SVM Experiments

To optimize the SVM classifier the γ value was varied from 10 to 100 in steps of 10 for each d value, which was also varied from 1 to 10 in steps of 1. The input features were also varied from 1 to 3 and for each of these 100 configurations, five experiments were carried out using the Jack-knife technique and the average *TPR* and *FPR* values were recorded. After these experiments the average *TPR* and *FPR* values, obtained for 300 SVM configurations, were compared using the ROC curve of Figure 5.4.

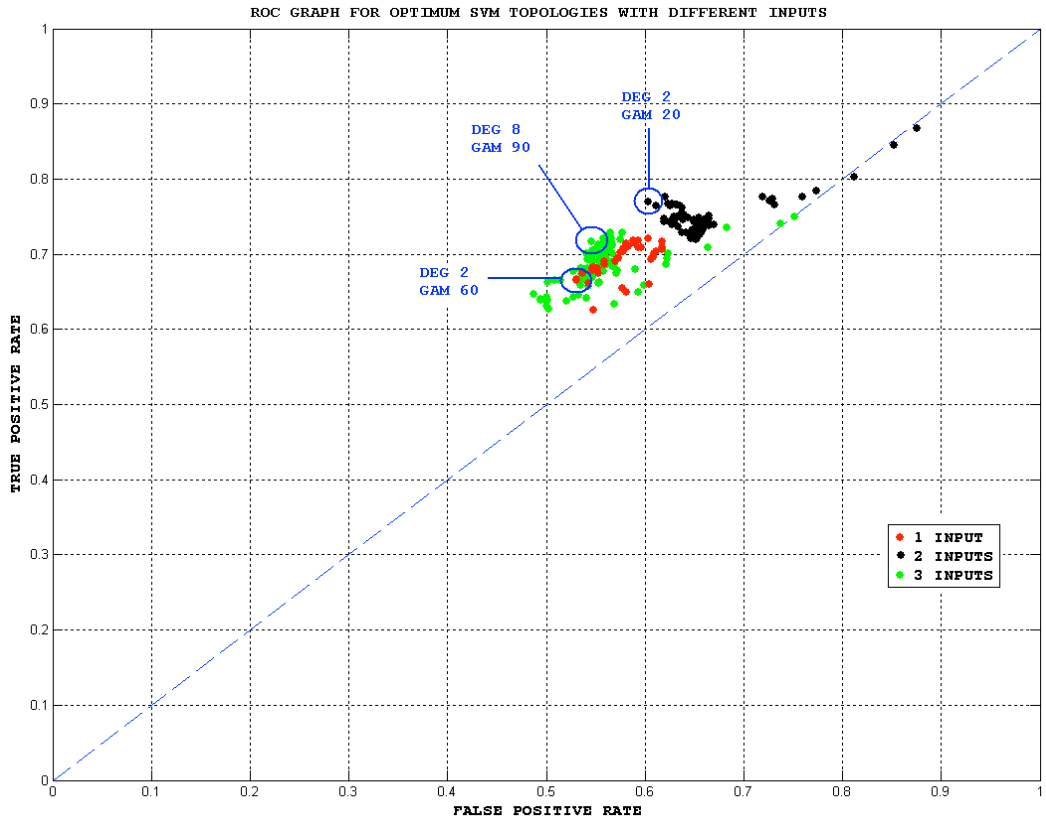


Figure 5.4 ROC graph showing the best SVM topologies with different inputs.

In order to find the classification thresholds that provide the best predictions for the optimum SVM topologies, the threshold values were changed from 0 to 1 in steps of 0.01 for every input and their selected optimum topologies. Then for each threshold value, five experiments were carried out using the Jack-knife technique and the average *TPR* and *FPR* values were found and an ROC curve was created as shown in Figure 5.5.

At the end of these experiments it was concluded that an SVM classifier that accepts three inputs with d and γ values set to 8 and 90, respectively and a classification threshold value of 0.83 provided the best prediction performance. This SVM configuration provided *TPR* and *FPR* of 0.73 and 0.53 respectively.

5.4.4 Further Experiments

It was decided to conduct further learning experiments with the classifiers to improve the prediction performance. To reduce the number of falsely associated CMEs, the association rules were modified by exploring other features provided in the CME

catalogue. As explained in Chapter 3, the Measurement Position Angle (MPA) can be used as an indicator of the CPA of a halo CME. Hence, MPA in the CME catalogue was used in further experiments to provide indications of the locations of associated flares. The rules of association between CMEs and flares, which are explained in Chapter 4, have been modified to include MPA as a second criterion of comparison besides timing. Applying this extra feature has reduced the number of associated flares. The new set, obtained using parameter values α = 150 minutes and β = 60 minutes, consists of 405 *A* flares and 404 *NA* flares.

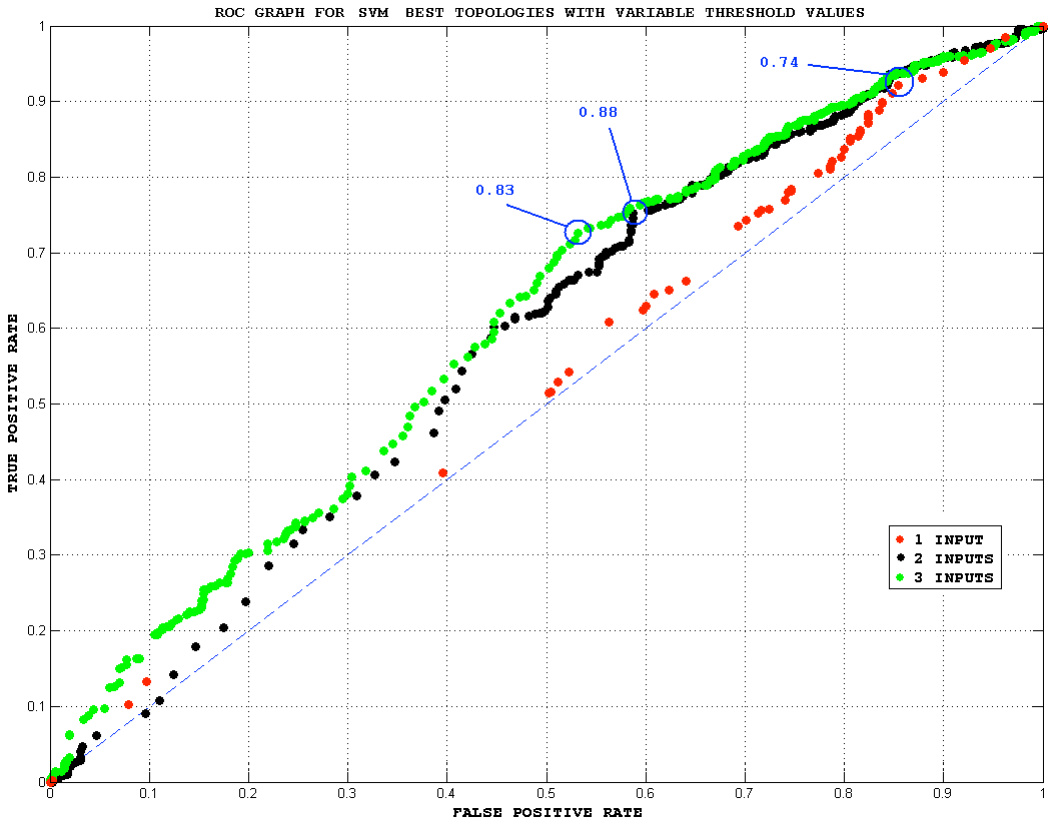


Figure 5.5 ROC graph showing the best SVM topologies with different inputs variable threshold values.

The optimisation and learning experiments using CCNNs and SVMs were carried out as explained in the previous two subsections. At the end of these experiments the optimum configuration obtained for a SVM with 3 inputs was 8, 90 and 0.72 for *d*, γ and classification threshold, respectively. This configuration provides *TPR* and *FPR* rates of 0.74 and 0.59 respectively. On the other hand the optimum topology

for a CCNN is 3 input nodes, with 3 hidden nodes and a classification threshold of 0.47. This topology generates *TPR* and *FPR* values of 0.71 and 0.46 respectively.

It is clear from these results that the prediction performances for both CCNNs and SVMs have been improved and the *FPR* has been reduced. The use of MPA for a location-based association enabled the association sets to be refined and hence eliminated some of the false associations, which produced some improvement in the prediction performance.

5.5 CME Predictions Based on CME-Filament Associations

5.5.1 Data Handling

The associations between CMEs and filaments were investigated in Chapter 4 for two groups of data. These associations were processed as described here using different machine learning algorithms for the purpose of CME predictions. SVMs (Al-Omari et al., 2008) and RBFs (Qahwaji et al., 2008a) are used to analyse associations of group 1 data. The data of group 2 is processed using the AdaBoost algorithms (Qahwaji et al., 2008b) and more improvements are done using SVMs (Al-Omari et al., 2009a).

Numerical representations were used for *A* and *NA* filaments as machine learning algorithms deal mainly with numbers. Properties such as starting time, ending time, type and spatial extent of the filaments can be extracted from the NGDC filaments catalogue. Initially the inclusion of other properties such as filament location, orientation and importance were considered, but unfortunately the necessary data are not provided for a large proportion of the associated filaments and the only location indicator that is available for all filaments is the centroid location. For example, about 63% (4606 out of 7332) of the filament records of group 2 data, in the period 1996-2001, are reported without importance. Hence, it was decided to use only the groups of properties shown in Table 5-2.

Table 5-2 Groups of properties that are used as input nodes in the SVM learning algorithm.

Group	Inputs
4	Timing, duration, type, extent
3	Timing, duration, type
3a	Timing, duration, extent
3b	Timing, type, extent
2	Timing, duration
2a	Timing, type
2b	Timing, extent

The timings in Table 5-2 represent Julian dates of the filaments. As explained before, for group 1 the filament event time (used for time-based association) was considered to be the filament end time. And for group 2 the filament event time was considered to be the average of the filament start and end times. Values of the Julian date within solar cycle 23 are of the order of ~ 2,450,000, with an increment of 1 each day. So, the Julian date was calculated and normalised to be in the range between 0.1 and 0.9. Filament distribution according to the solar cycle time is shown in Figure 5.6 for both *A* and *NA* filaments in data of group 2.

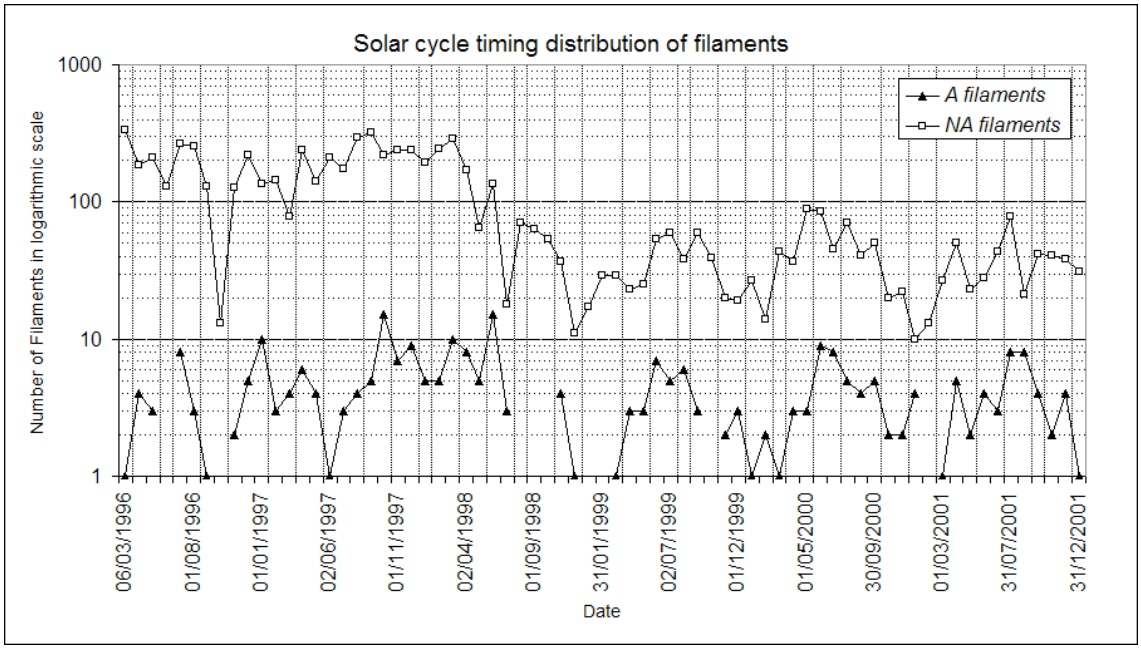


Figure 5.6 Solar cycle timing distribution for CME-associated and not-associated filaments within data group 2.

The filament duration was calculated as the time difference in hours between the end and start times and then it was normalised between 0.1 and 0.9. The duration distributions for *A* and *NA* filaments are shown in Figure 5.7 for the data of group 2. In the same manner, the filament extent was normalized in the range from 0.1 to 0.9 and its distribution for group 2 data is shown in Figure 5.8.

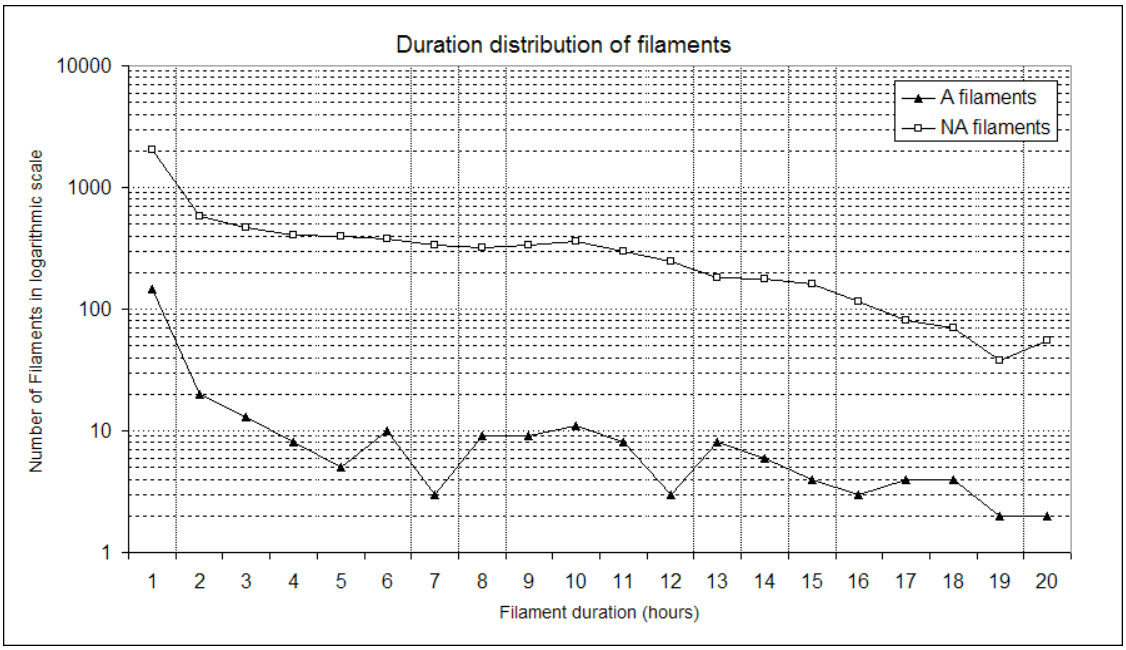


Figure 5.7 Duration distributions for CME-associated and not-associated filaments within data group 2.

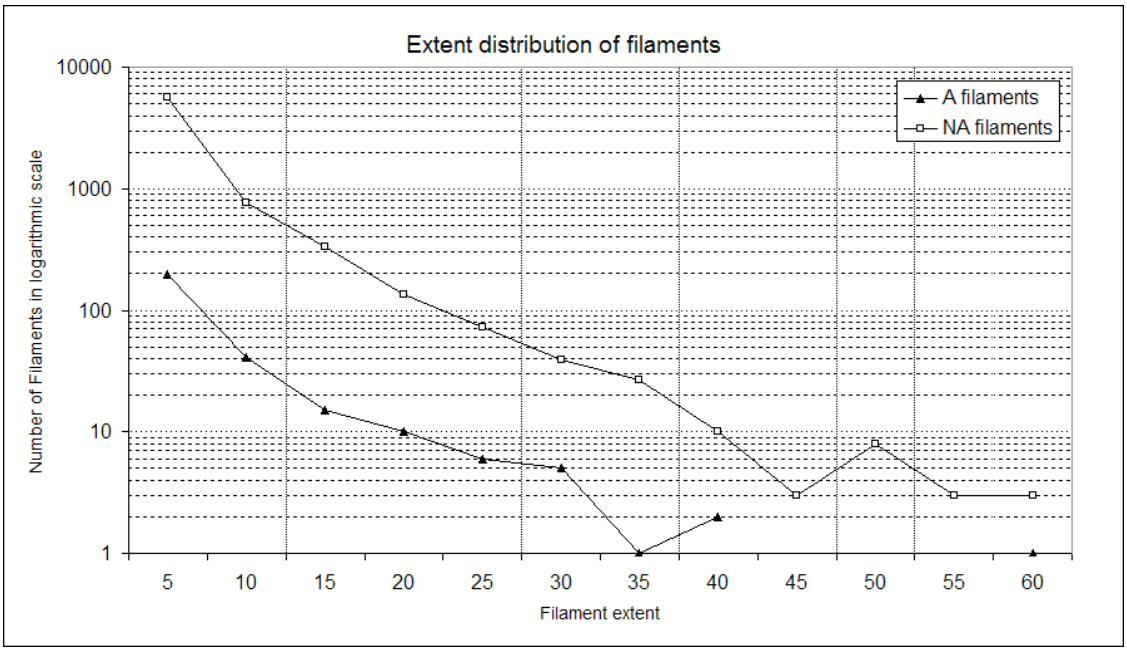


Figure 5.8 Extent distributions for CME-associated and not-associated filaments within data group 2.

For the filament type parameter to have a meaningful numerical value it can be represented by its probability within the associated filaments and this probability can be calculated from the distribution of filament types of Figure 5.9.

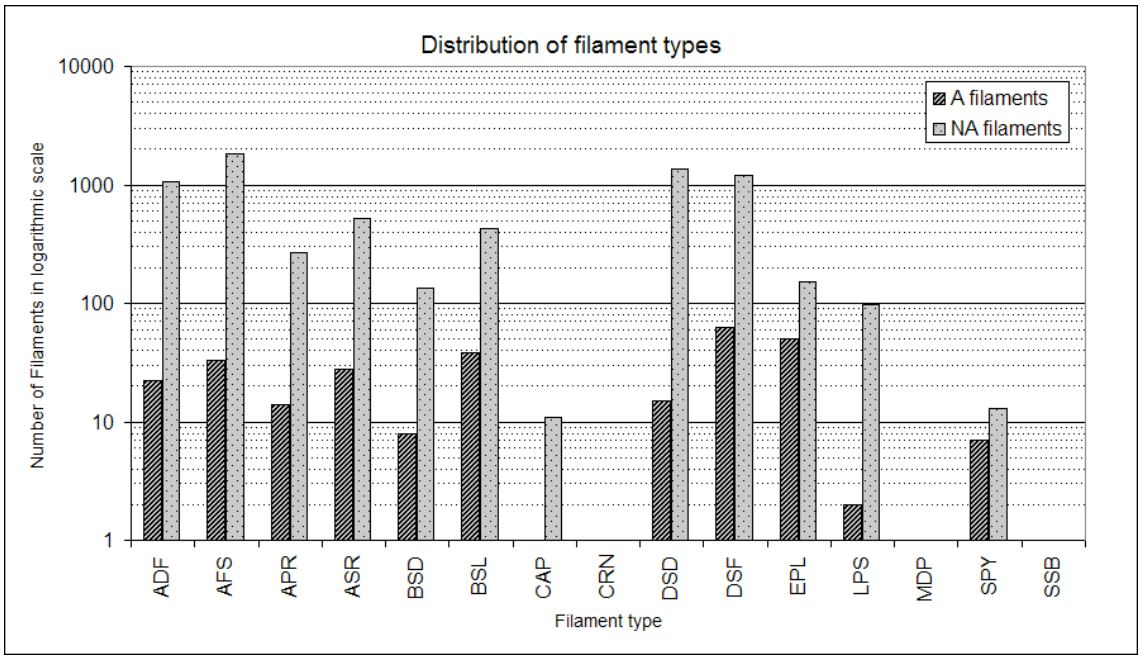


Figure 5.9 Type distributions for CME-associated and not-associated filaments within data group 2.

However, some types have almost equal numbers of associated filaments such as the DSD and APR events while other types such as CAP, CRN, MDP, and SSB are not associated with CMEs. In such cases, it will be impossible for the SVM classifier to distinguish between different types of filaments because they are represented by values that are not separated enough for successful learning and output class separation. Hence, it has been decided to represent the filament classes in numerical codes for the learning experiments as listed in Table 5-3. It is important to point out that these numerical values are no more than codes assigned to each class; they are neither weights nor represent the probability distribution of these classes.

Finally, the target function for the input groups is represented by two values: 0.9 indicates that the filament is initiating a CME and 0.1 indicates that the filament is not initiating a CME.

Table 5-3 Numerical representation for the filament types.

Type	Numerical Value	
	Group 1	Group 2
SSB	0.10	0.10
MDP	0.15	0.15
CRN	0.20	0.20
CAP	0.25	0.25
LPS	0.40	0.30
SPY	0.30	0.35
BSD	0.35	0.40
APR	0.50	0.45
DSD	0.75	0.50
ADF	0.70	0.55
ASR	0.60	0.60
AFS	0.85	0.70
BSL	0.55	0.75
EPL	0.45	0.85
DSF	0.90	0.90

5.5.2 CME Predictions using data of group 1

After creating the associated data set described in the previous chapter, the training and testing experiments for the machine learning algorithms were carried out. For the current group of data, SVMs and RBFs are optimised and compared in the context of CMEs prediction. All experiments were carried out with the aid of the Jack-knife technique and the prediction performance was evaluated using ROC curves based on two performance indicators: *TPR* and *FPR*. A total of 2221 associated and not-associated filaments were used for training. This constituted 80% of the total number of associated cases available. The remaining 555 associated and not-associated filaments were used for testing. The above numbers are true for all the input groups that have no features representing extension. Unfortunately, the extensions are not always indicated in the filament catalogues. Due to this lack of data, the number of associated sets was reduced to 1966 for groups 2b, 3b, 3a and 4.

5.5.2.1 SVM Experiments

The Anova-Kernel SVM, explained in Section 5.2, was used in all the SVM experiments. To optimize the SVM classifier the γ value was varied from 1 to 10 in steps of 1 for each value of d, which was also varied from 1 to 10 in steps of 1. The input features were also varied in the seven groups shown in Table 5-2. For each of these 100 configurations, 10 experiments were carried out using the Jack-knife technique and the average *TPR* and *FPR* values recorded. Hence, 700 experiments were carried out with 700 SVM configurations, resulting in 70 average *TPR* and *FPR* values being produced. These values are plotted in Figure 5.10 to find the optimum degree and gamma values and optimum inputs configuration.

As explained previously, the best performance shown in ROC curves is the one furthermost from the random guessing diagonal line. So, the optimum SVM configurations were found as shown in Figure 5.11 which is the magnified region labelled Z in Figure 5.10.

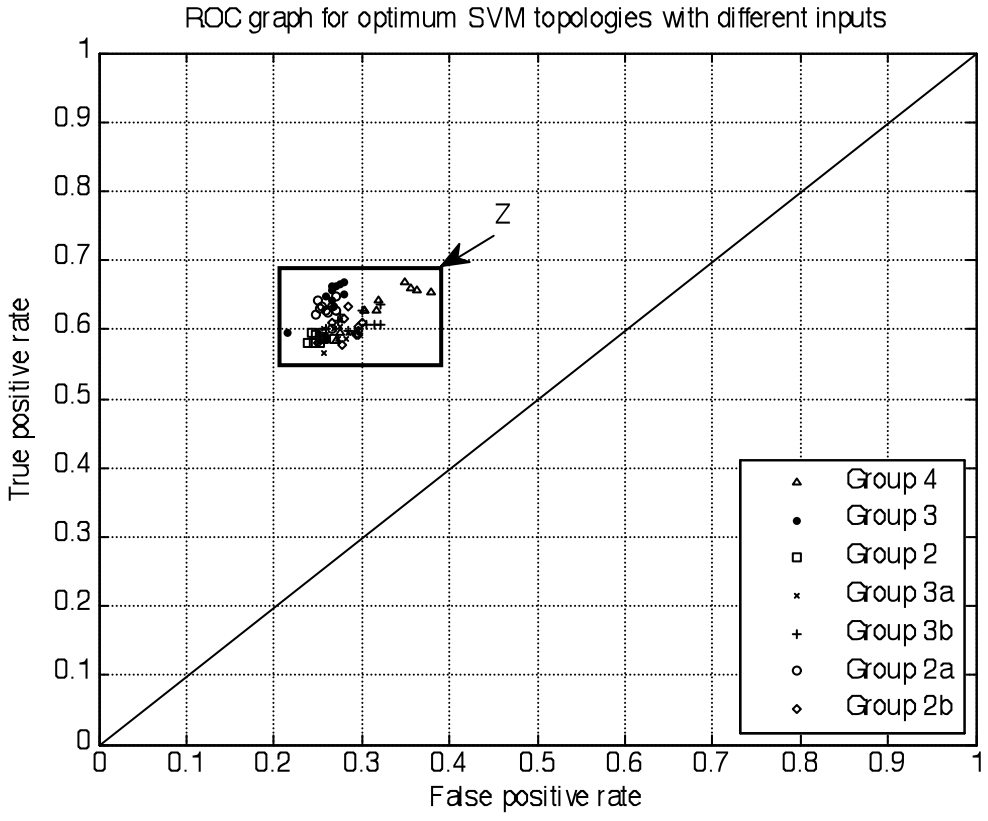


Figure 5.10 ROC graph showing different SVM topologies with different inputs.

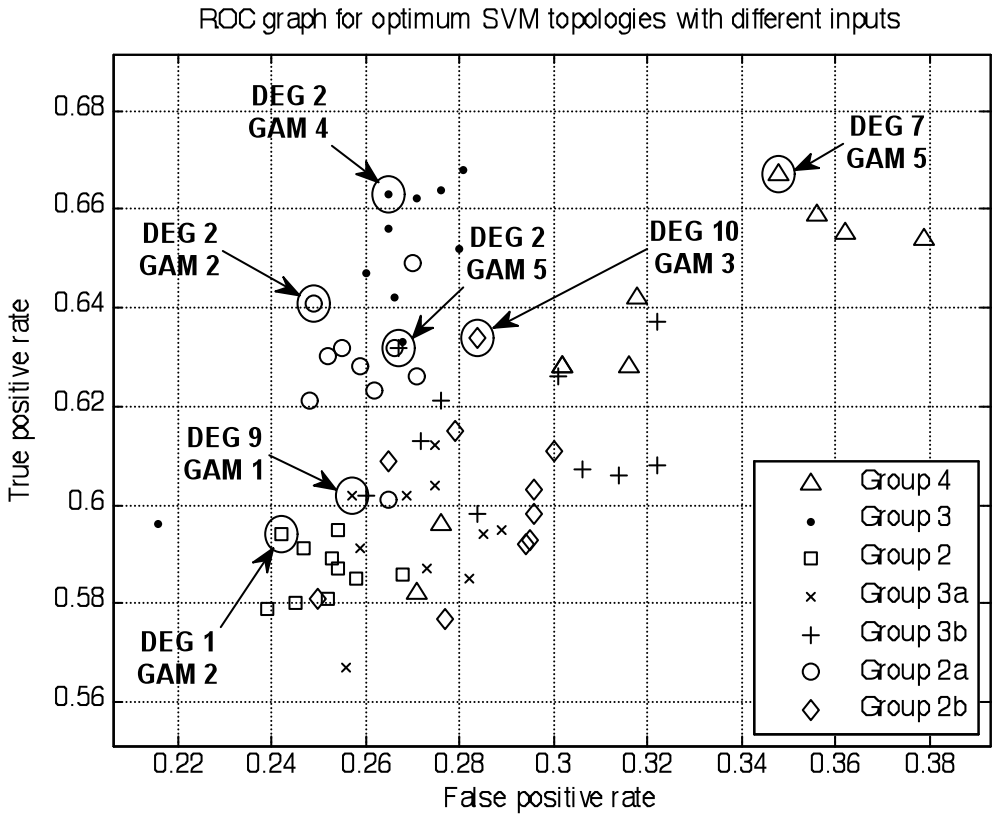


Figure 5.11 Magnified box Z in Figure 5.10: ROC graph showing the optimum SVM topologies with different inputs.

In order to find the classification thresholds that provide the best predictions for the optimum SVM topologies, the threshold values were changed from 0 to 1 in steps of 0.01 for every input and their selected optimum topologies. Then for each threshold value, 10 experiments were carried out using the Jack-knife technique and the average *TPR* and *FPR* values were calculated. At the end of these experiments the resulting ROC curve is shown in Figure 5.12.

The optimum threshold values are found by choosing the threshold value with performance closest to the northwest corner in the ROC curve. This is shown clearly by magnifying region Z of Figure 5.12 as depicted in Figure 5.13.

As can be seen from figures 5.11 and 5.13, an SVM classifier that accepts two inputs (group 2a) with degree and gamma values of 10 and 3 respectively and a classification threshold value of 0.74 provides the best prediction performance. This

SVM configuration provides average *TPR* and *FPR* of 0.640 and 0.254, respectively (Al-Omari et al., 2008).

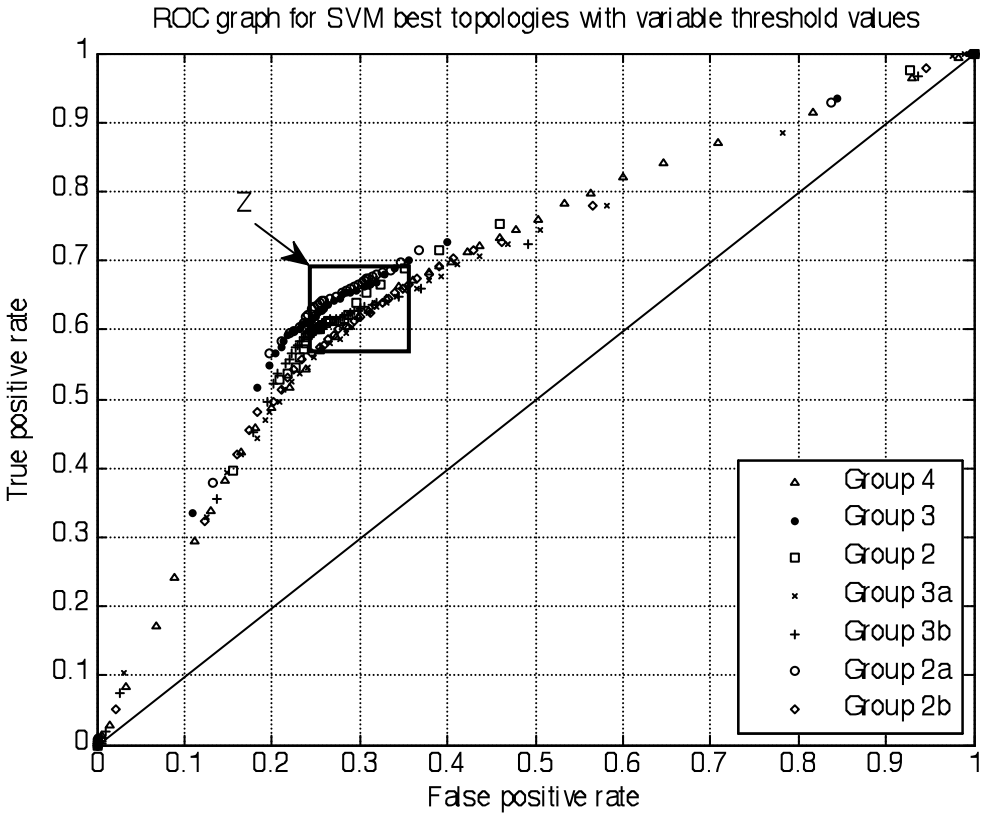


Figure 5.12 ROC graph showing different SVM topologies with variable threshold values.

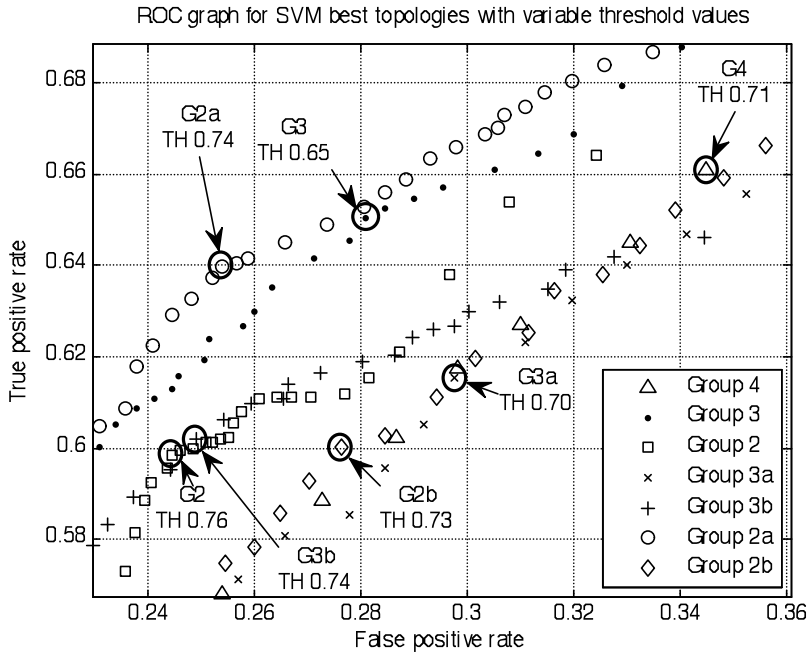


Figure 5.13 Magnified box Z in Figure 5.12: ROC graph showing the best SVM topologies with variable threshold values.

The next best performance is achieved by using three inputs (group 3) with degree, gamma and threshold values of 2, 4 and 0.65, respectively. This SVM configuration provides *TPR* and *FPR* of 0.651 and 0.281 respectively.

In general, there has been an increase in the prediction rate with the use of more discriminative input features, such as the filament type, compared to the input groups of Table 5-2. These experiments indicate that the filament type and duration are more important for CME prediction than the filament extent.

5.5.2.2 RBF Experiments

Optimisation of the learning algorithm is usually carried out to find the best parameters and/or topologies for the learning algorithms. For this work, RBFs with high spread values to ensure accurate fitting, were used to ensure that optimum RBF performance can be obtained.

Intensive training totalling 100 experiments were carried out for every group of input features. Each experiment consisted of training a random set of the associated sets followed by a testing phase on the remaining sets. As explained previously these sets were obtained using the Jack-Knife technique and were quite different for every experiment. The ROC performance indicators were found for every experiment. These indicators are *TPR*, *FPR*, *FNR*, *TNR*, accuracy, specificity and sensitivity, all calculated as explained in Section 5.3. Averages of these indicators for the 100 experiments were found for every group and are shown in Table 5-4. A total of 700 experiments were performed in finding these values.

The average *TPR* and *FPR* values are plotted in Figure 5.14 and are used to determine the input group providing the best performance. It is clear from Figure 5.15 that input group 3 provides the best performance with an accuracy of 69% followed by group 2a.

Table 5-4 Average ROC performance indicators for different input combinations.

Group	*TPR*	*FPR*	*FNR*	*TNR*	Accuracy	Specificity	Sensitivity
4	0.604	0.273	0.396	0.727	0.663	0.727	0.604
3	0.65	0.269	0.35	0.731	0.69	0.731	0.65
2	0.615	0.275	0.385	0.725	0.67	0.725	0.615
3a	0.58	0.254	0.42	0.746	0.66	0.746	0.58
3b	0.609	0.281	0.391	0.719	0.662	0.719	0.609
2a	0.65	0.296	0.35	0.704	0.676	0.704	0.65
2b	0.593	0.268	0.407	0.732	0.659	0.732	0.593

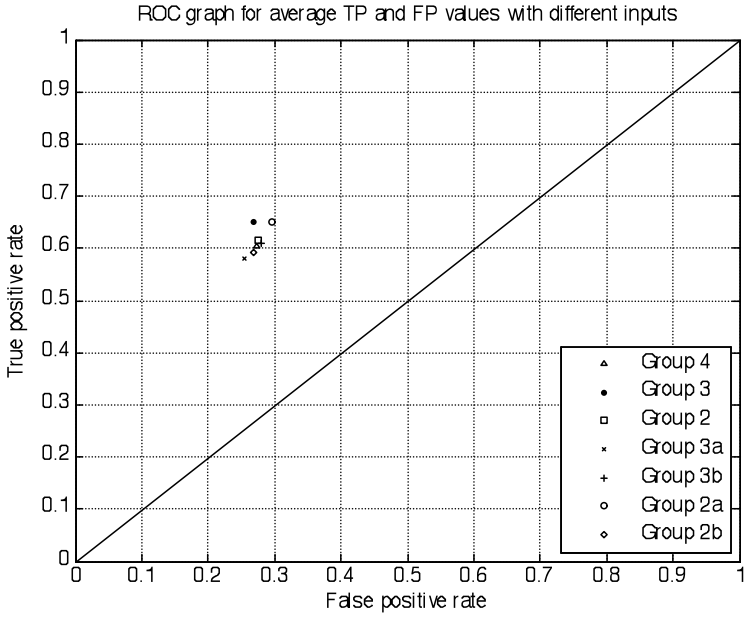


Figure 5.14 ROC graph showing the average *TPR* and *FPR* values.

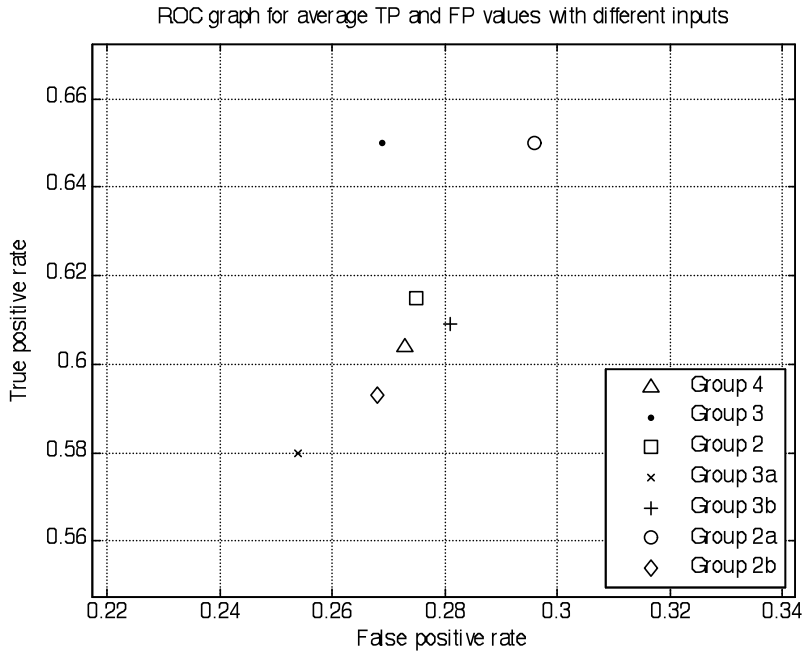


Figure 5.15 Magnification of Figure 5.14 ROC graph showing the average *TPR* and *FPR* values.

5.5.3 CME Predictions using data of group 2

In the previous subsection it was found that group 3 and group 2a were the best input groups for the purpose of CME predictions (Al-Omari et al., 2008, Qahwaji et al., 2008a). Nevertheless, another extensive set of experiments was carried out as described here, attempting to increase the accuracy of the proposed prediction system and to determine the significance of each property within this context.

Training datasets were created including 40% *A* filaments and 60% *NA* filaments. Training and testing experiments were carried out for this group of data using boosting algorithms and SVMs. The prediction performances for SVMs were evaluated using two validation methods as explained below.

5.5.3.1 AdaBoost Algorithm Experiments

By applying the association algorithm for group 2 (Chapter 4) an associated data set, consisting of 522 filaments with 209 (40%) *A* filaments and 313 (60%) *NA* filaments, was created. For this work, three different boosting algorithms were used: Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. All the machine learning/training and testing experiments were carried out with the aid of the Jack-knife technique.

Intensive training totalling five learning experiments, with 1000 iterations for each experiment, were carried out for the optimum group of input features, which was found to be group 3 in the previous subsection. Each experiment involved training on a random set of associated sets followed by a testing phase on the remaining sets. The ROC performance indicators were found for every experiment. These indicators are *TPR*, *FPR*, *FNR*, *TNR*, accuracy, specificity and sensitivity, all calculated as explained in Section 5.3. The averages of these indicators over all the experiments and for the three AdaBoost algorithms are shown in Table 5-5.

Table 5-5 Average ROC performance indicators for different input combinations.

	AdaBoost Algorithm		
	Real	Gentle	Modest
TPR	0.57407	0.46296	0.57407
FPR	0.25658	0.11842	0.26316
FNR	0.42593	0.53704	0.42593
TNR	0.74342	0.88158	0.73684
Accuracy	0.67308	0.70769	0.66923
Specificity	0.74342	0.88158	0.73684
Sensitivity	0.57407	0.46296	0.57407
HSS	0.31968	0.36649	0.31253

The *TPR* and *FPR* values obtained from all the experiments on the three types of AdaBoost algorithm are plotted in Figure 5.16 and used to determine the input group providing the best performance.

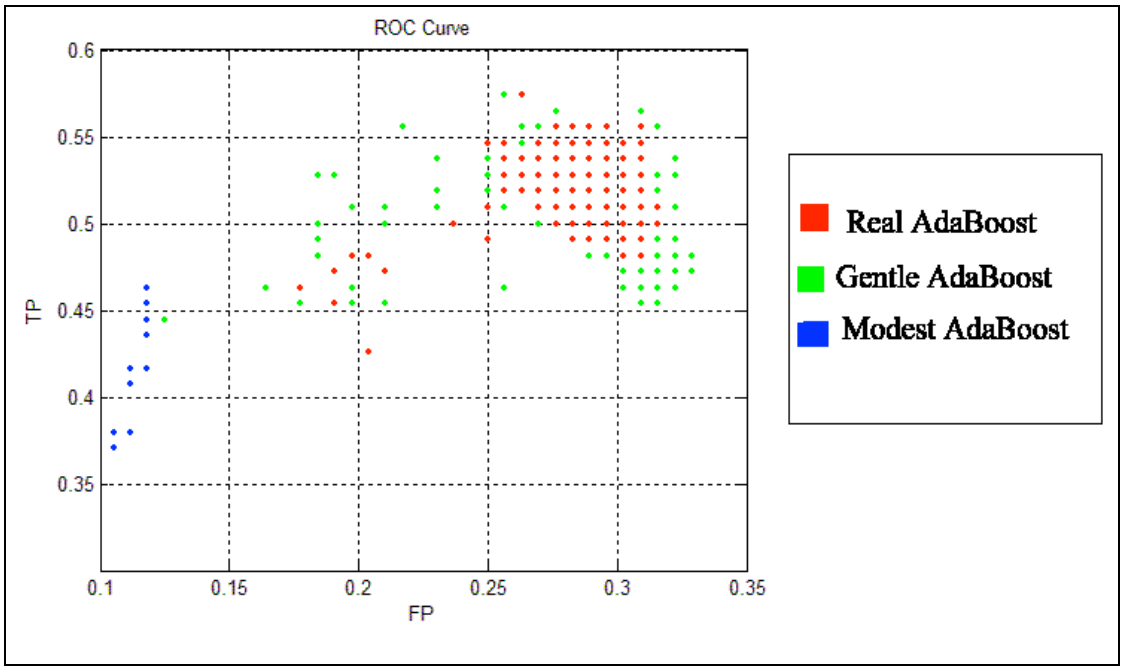


Figure 5.16 The ROC graph for the AdaBoost learning experiments.

As shown in Table 5-5 the Adaboost algorithms provide good accuracy compared to the performance of Radial Basis Functions (Qahwaji et al., 2008a), but still not very high. Compared to Radial Basis Functions they provide lower *TPR* but also higher *TNR*. The best prediction performance, measured in terms of the accuracy of predictions is 70.8% and is provided by the Gentle AdaBoost. However, the high accuracy value is caused mainly by the high *TNR* value which is 88.1% not by the *TPR*

value which is 46.3%. This *TPR* value is very low and even random guessing would perform better than this. On the other hand, the gentle AdaBoos provides the best *HSS* value which is ~0.37. Hence, a major conclusion of these results is that the Gentle AdaBoost can provide reliable performance if used as a rejection classifier to predict when CMEs are not likely to occur. It is less effective if used as a positive classifier tool.

The Real and Modest AdaBoosts provide higher *TPR* rates and acceptable *HSS* values but their accuracy and *TNR* rates are lower. It is worth noting that the *TPR* rates provided by the Real and Modest AdaBoosts are not as high as the *TPR* rates provided by the RBFNs and SVMs used in the previous subsection (Qahwaji et al., 2008a). Hence, it is not very appropriate to use them within the context of this problem.

5.5.3.2 SVM Experiments - Validation Method 1

As explained before, the performance of the Anova-Kernel SVM is optimised by adjusting the values of d, γ, and classification threshold. In the optimisation process, the γ and d values were both varied from 1 to 10 in steps of 1. In all experiments, the classifier threshold was initialized to the mean of the predicted scores. The optimisation process was applied to input features corresponding to each of the seven groups shown in Table 5-2.

The machine learning/training and testing experiments in validation method 1 were carried out with the aid of the Jack-knife technique. As mentioned previously, the learning dataset contains 209 *A* filaments representing 40% of the dataset and another 313 *NA* filaments (60%) were selected randomly to build a complete dataset of 522 filaments. A total of 418 associated and not-associated filaments were used for training. This constituted 80% of the total number of cases. The remaining 104 associated and not-associated filaments were used for testing. Again, these numbers apply to all the input groups that have no features representing extension (groups 3, 2 and 2a).

Unfortunately, filament extensions are not always reported in the NGDC catalogues so the associated filaments, reported without extent, were discarded from the training and testing datasets in groups 4, 3a, 3b and 2b. In these cases, the number of *A* filaments is reduced to 143, which means there are only 214 *NA* filaments and a total of 357 associated and not-associated filaments. Hence, for groups 4, 3a, 3b and 2b we have a dataset of 287 associated and not-associated filaments for training and another set with the remaining 70 associated and not-associated filaments for testing.

For each of 100 configurations and seven input groups, 10 experiments were carried out using the Jack-knife technique and the average *TPR* and *FPR* values recorded. Hence, 7000 experiments were carried out with 700 SVM configurations to produce 700 average values of *TPR* and *FPR*. To find the optimum SVM system (optimum d, γ and input configuration), the results were analysed using the ROC analysis technique and are plotted in Figure 5.17. The best performing SVM configurations can be seen in Figure 5.18, which is the magnified region labelled Z in Figure 5.17.

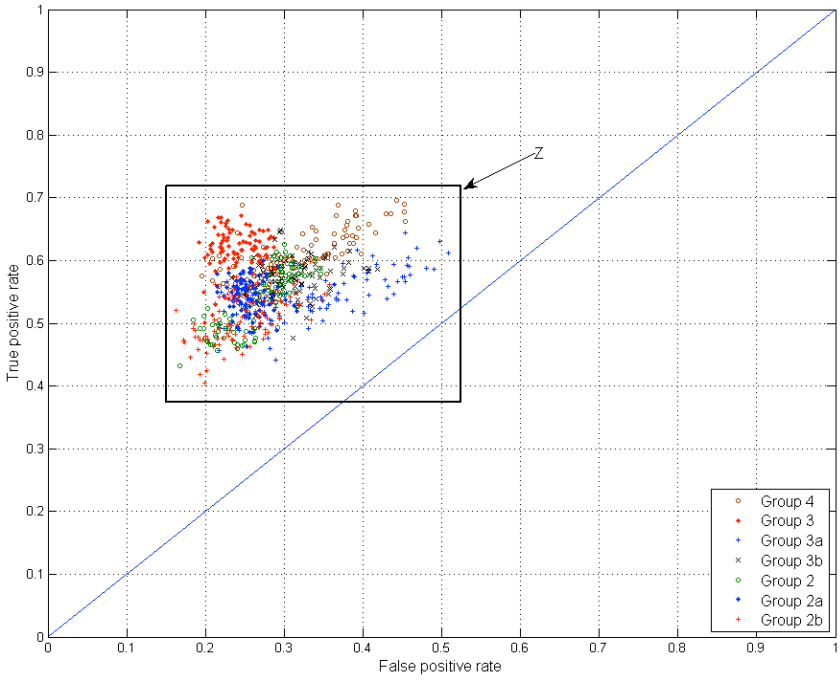


Figure 5.17 ROC graph showing different SVM topologies with variable *d* and *γ* values for validation method 1.

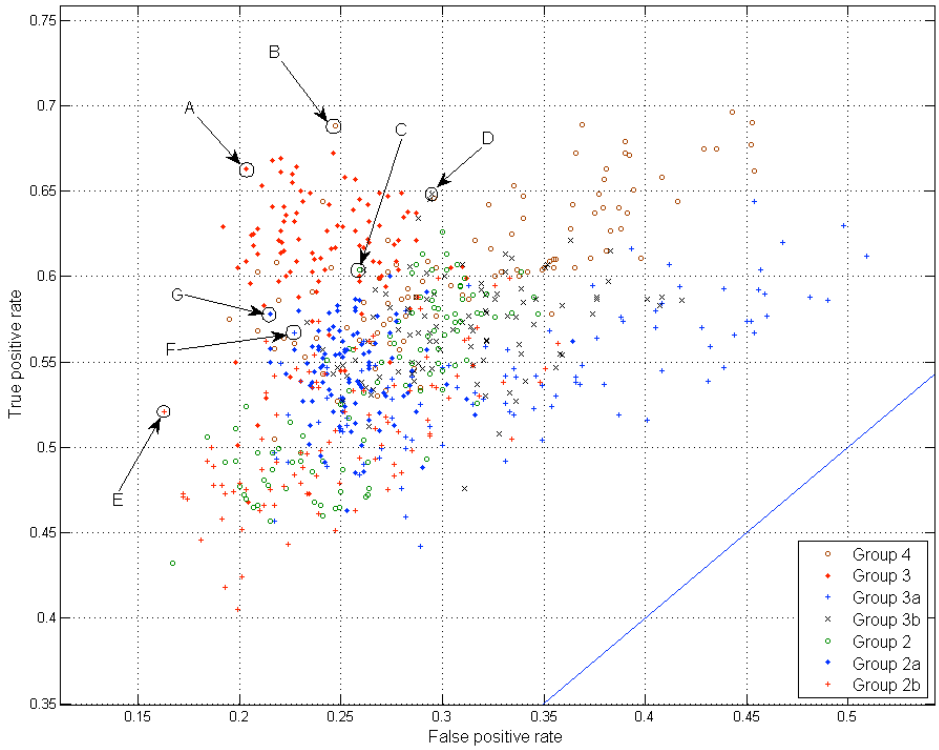


Figure 5.18 Magnified view of region Z in Figure 5.17: ROC graph showing the optimum SVM topologies with variable *d* and *γ* values for validation method 1. The (*d*, *γ*) values for the optimum topologies are: A(2,8), B(1,6), C(7,8), D(3,8), E(2,9), F(8,1), G(10,7).

In order to find the classification thresholds that provide the best predictions for the optimum SVM topologies, the threshold values were changed from 0 to 1 in steps of 0.01 for every input feature set and their selected optimum topologies. Then for each threshold value, 10 experiments were carried out using the Jack-knife technique and the averages for all performance indicators, defined previously, were calculated. The results of these experiments are summarized in Table 5-6 and depicted in the ROC curve of Figure 5.19.

Table 5-6 Averages of performance indicators (Jack-knife technique)

Group	*d*	*γ*	*TPR*	*FPR*	*FNR*	*TNR*	Accuracy	Specificity	Sensitivity	*HSS*	D_{ROC}	Threshold
4	1	6	0.60	0.25	0.40	0.75	0.69	0.75	0.60	0.35	0.250	0.56
3	2	8	0.65	0.22	0.35	0.78	0.73	0.78	0.65	0.43	0.304	0.57
3a	8	1	0.60	0.27	0.40	0.73	0.67	0.73	0.60	0.33	0.234	0.51
3b	3	8	0.61	0.29	0.39	0.71	0.67	0.71	0.61	0.31	0.229	0.53
2	7	8	0.67	0.36	0.33	0.64	0.65	0.64	0.67	0.30	0.219	0.55
2a	10	7	0.62	0.24	0.38	0.76	0.70	0.76	0.62	0.38	0.269	0.55
2b	2	9	0.59	0.29	0.41	0.71	0.66	0.71	0.59	0.29	0.209	0.49

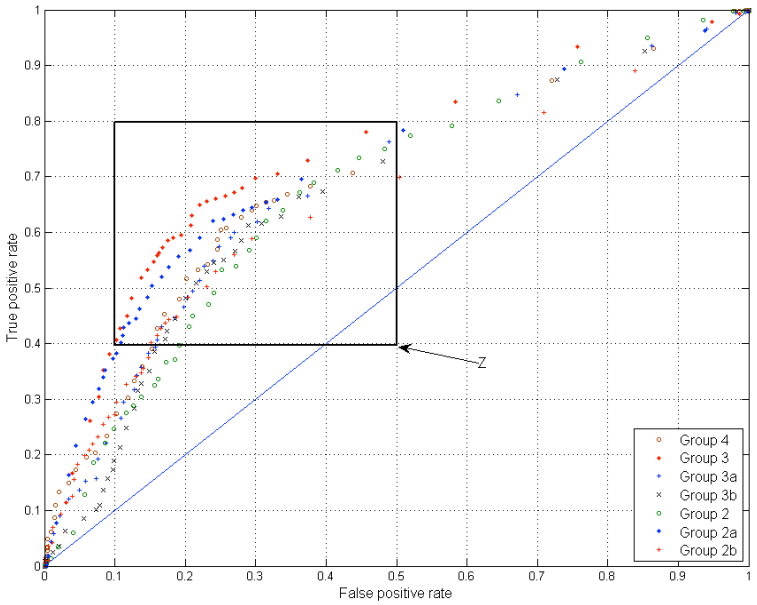


Figure 5.19 ROC graph showing different SVM topologies with variable threshold values for validation method 1.

The optimum threshold values were found by choosing the threshold value with the system performance closest to the upper-left corner in the ROC curve. This is seen clearly in Figure 5.20, which shows a magnified view of the region labelled Z in Figure 5.19.

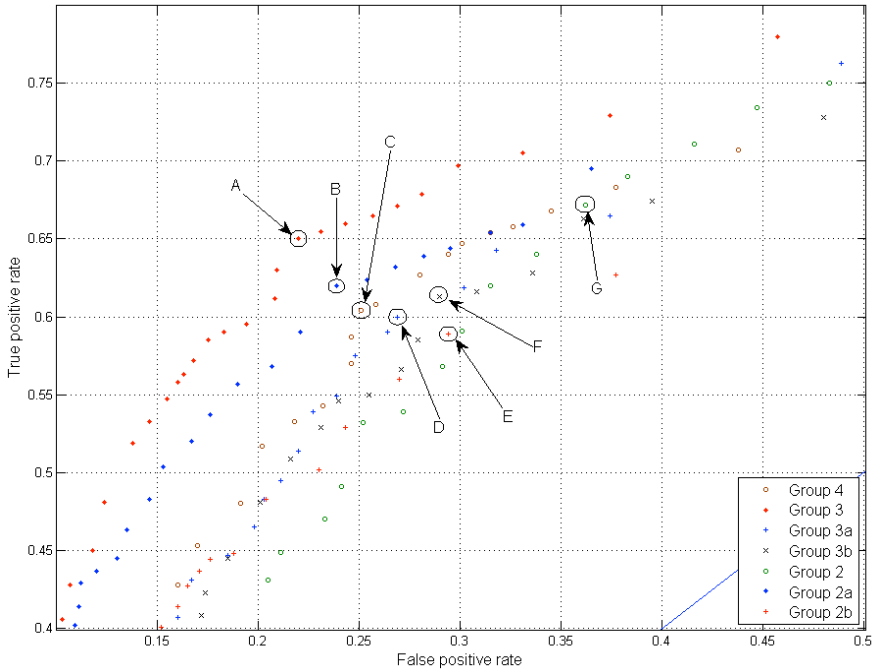


Figure 5.20 Magnified view of region Z in Figure 5.19: ROC graph showing the best SVM topologies with variable threshold values for validation method 1. The threshold values for the optimum topologies are: A(0.57), B(0.55), C(0.56), D(0.51), E(0.49), F(0.53), G(0.55).

As can be seen by inspection of Figure 5.18 and Figure 5.20, an SVM classifier that accepts three inputs (group 3) with d and γ values of 2 and 8 respectively and a classification threshold value of 0.57 provides the best prediction performance. It achieved average *TPR*, *FPR*, and *TNR* values of 0.65, 0.22, and 0.78, respectively. This is a good result as it corresponds to an average accuracy of 73% and a Heidke skill score of 0.43.

The next best performance is achieved by using two inputs (group 2a) with d, γ and threshold values of 10, 7 and 0.55, respectively. This SVM configuration provides *TPR*, *FPR*, specificity, accuracy and *HSS* of 0.62, 0.24, 76%, 70% and 0.38, respectively.

To draw an accurate conclusion on the importance of filament properties in CME prediction, the same dataset size must be used during validation. So, further experiments were carried out using the same datasets used before for input groups 4, 3a and 3b except that the extent property was discarded from these datasets. For comparison purposes, the groups were relabelled as 4′, 3a′ and 3b′. Validation method 1 was used and the optimum results of the experiments are summarized in Table 5-7.

Table 5-7 Averages of performance indicators (discarding extent from inputs).

Group	d	γ	*TPR*	*FPR*	*FNR*	*TNR*	Accuracy	Specificity	Sensitivity	*HSS*	D_{ROC}	Threshold
4′	1	6	0.66	0.26	0.34	0.74	0.70	0.74	0.66	0.39	0.283	0.49
3a′	3	2	0.52	0.24	0.48	0.76	0.67	0.76	0.52	0.29	0.198	0.50
3b′	1	4	0.62	0.27	0.38	0.73	0.69	0.73	0.62	0.34	0.248	0.54

By comparing the values *TPR*, *FPR*, accuracy and *HSS* of group 4 in Table 5-6 with those of group 4′ in Table 5-7 it is clear that discarding the filament extent from the inputs enhanced the prediction performance. By doing the same comparison between the optimum results of groups 3a and 3b in Table 5-6 and groups 3a′ and 3b′ in Table 5-7 we can conclude that filament type and duration, particularly the former, are more important indicators for CME prediction than filament extent. This conclusion supports the findings of some researchers who reported high associations between CMEs and

certain types of filaments. An example of this is the study reported by Pojoga and Huang (2003) where the authors considered three classes of sudden disappearances: eruptive, quasi-eruptive and vanishing (thermal disappearances) filaments. They found that 70% of the eruptive filaments were associated with CMEs, while the correlations were weaker for quasi-eruptive and vanishing filaments.

5.5.3.3 SVM Experiments - Validation Method 2

The second validation method came in an attempt to measure the ability of the system design to constitute a near real-time automated CME prediction system. So, it was decided to validate the system on some arbitrary selected years of data without the need for random sampling of data using the Jack-knife technique. In this work extensive experiments were carried out using six years of data from 1996 to 2001. Here the data from years 1996, 1997, 2000 and 2001 were used for training and the years 1998 and 1999 were used for testing. A training dataset consisting of 149 *A* filaments and 223 *NA* filaments was created. The testing stage was more challenging because the testing dataset included all 1765 filaments reported in the NGDC catalogues for years 1998 and 1999. Again, because some filaments are reported without information on their spatial extent, the training and testing datasets were reduced while working with input groups 4, 3a, 3b and 2b. For training, a total of 265 filaments were used, consisting of 106 *A* and 159 *NA* filaments. The number of filaments used for testing was reduced to 1504.

A total of 100 experiments were carried out for each input group and the values of *TPR* and *FPR* were used to create the ROC curve shown in Figure 5.21 from which the optimum SVM configurations were found. To achieve the best performance of the prediction system the value of the classifier threshold was varied from 0 to 1 in steps of 0.01. The *TPR* and *FPR* values for all thresholds and for all input groups were used to create the graph of Figure 5.22 and all the performance indicators were calculated and summarized in Table 5-8.

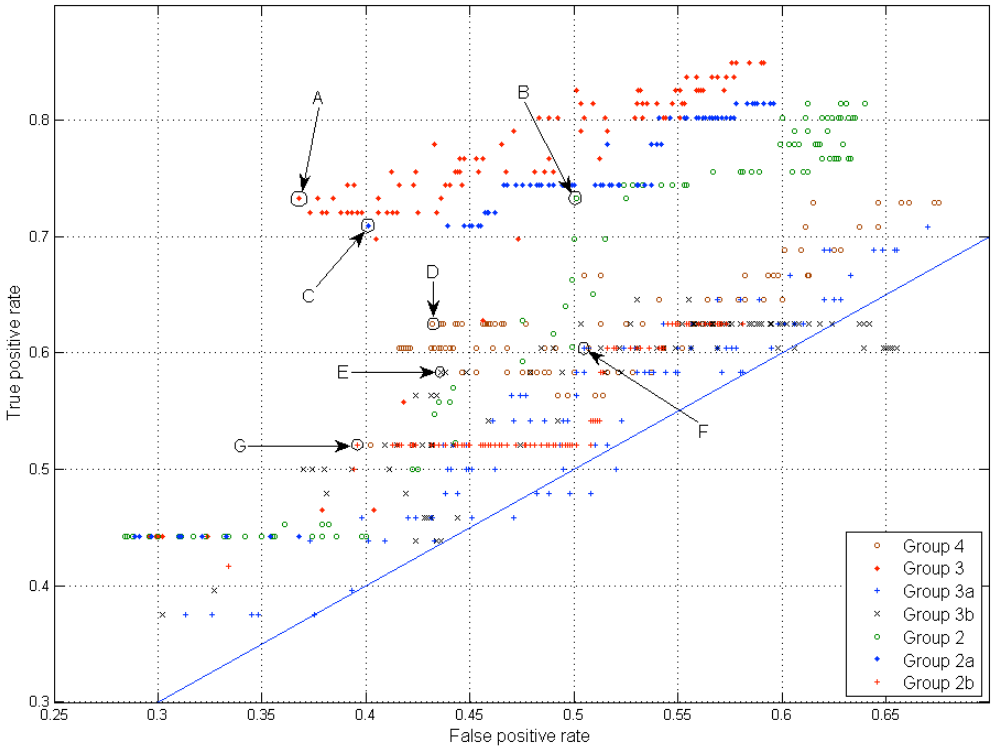


Figure 5.21 ROC graph showing the optimum SVM topologies with variable *d* and *γ* values for validation method 2. The (*d*, *γ*) values for the optimum topologies are: A(6,2), B(3,6), C(2,1), D(3,8), E(2,8), F(3,7), G(1,2).

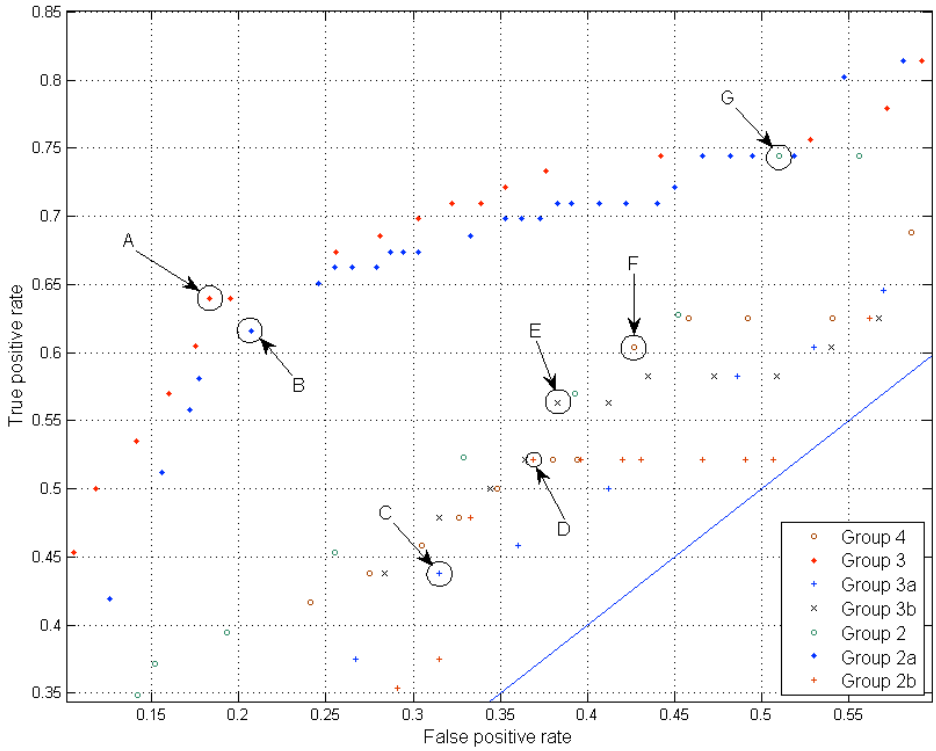


Figure 5.22 ROC graph showing the best SVM topologies with variable threshold values for validation method 2. The threshold values for the optimum topologies are: A(0.64), B(0.72), C(0.66), D(0.55), E(0.56), F(0.56), G(0.56).

Table 5-8 Averages of performance indicators (Validation method 2).

Group	d	γ	*TPR*	*FPR*	*FNR*	*TNR*	Accuracy	Specificity	Sensitivity	*HSS*	D_{ROC}	Threshold
4	3	8	0.60	0.43	0.40	0.57	0.57	0.57	0.60	0.03	0.126	0.56
3	6	2	0.64	0.18	0.36	0.82	0.81	0.82	0.64	0.18	0.323	0.64
3a	3	7	0.44	0.31	0.56	0.69	0.68	0.69	0.44	0.02	0.087	0.66
3b	2	8	0.56	0.38	0.44	0.62	0.62	0.62	0.56	0.03	0.127	0.56
2	3	6	0.74	0.51	0.26	0.49	0.53	0.49	0.74	0.04	0.166	0.56
2a	2	1	0.62	0.21	0.38	0.79	0.78	0.79	0.62	0.15	0.290	0.72
2b	1	2	0.52	0.37	0.48	0.63	0.63	0.63	0.52	0.02	0.108	0.55

From Figure 5.22 and Table 5-8 it is clear that the best performance was obtained while using group 3 with d, γ, and classification threshold values of 6, 2, and 0.64, respectively. This SVM configuration provides *TPR*, *FPR*, specificity, accuracy and *HSS* values of 0.64, 0.18, 82%, 81% and 0.18, respectively. So, the system uses SVM to predict if a CME is likely to be initiated with accuracy of 81% and at the same time to predict when CMEs are not likely to occur with specificity of 82%. Again, the next best performance was obtained with group 2a with d, γ, and classification threshold values of 2, 1, and 0.72, respectively. This configuration provides *TPR*, *FPR*, specificity, accuracy and *HSS* of 0.62, 0.21, 79%, 78% and 0.15, respectively.

5.6 Investigating the Associations among Sunspots, Flares and CMEs

As shown in the last two sections, the provided CME predictions are based on the associations between CMEs, filaments and flares. Hence, if an efficient flare prediction system exists in parallel with an automated filaments detection and classification system, then automated real-time CME predictions could be achieved. Automatic prediction of solar flares can be provided by applying machine learning algorithms to the associations between flares and sunspots. Such prediction system is reported in Qahwaji and Colak (2007) where sunspots and flares data for the period from 1 January 1992 to 31 December 2005 were processed to associate X- and M-class flares with their corresponding sunspots. It was concluded that SVMs provided the best performance for predicting if a sunspot group is going to flare or not. On the other hand,

CCNNs gave better results in predicting the class of the flare to erupt. Hence, a hybrid system combining SVM and CCNN was suggested.

Qahwaji and Colak (2007) considered M- and X-class flares only while C-class flares were not included in their study. In addition, the data used in their study consisted of mixed events from both solar cycles 22 and 23. Hence it was decided to extend their association algorithm to include C-class flares and to try to improve their system performance by including more data from solar cycle 23 and excluding any events within solar cycle 22. Furthermore, the associations between CMEs and the sunspot-associated flares were investigated for the purpose of CME predictions.

All the associations between sunspots and solar flares were found using the sunspot-flare association algorithm described in Chapter 4. Then, a numerical dataset was created and used in the learning mode as shown in Figure 5.23.

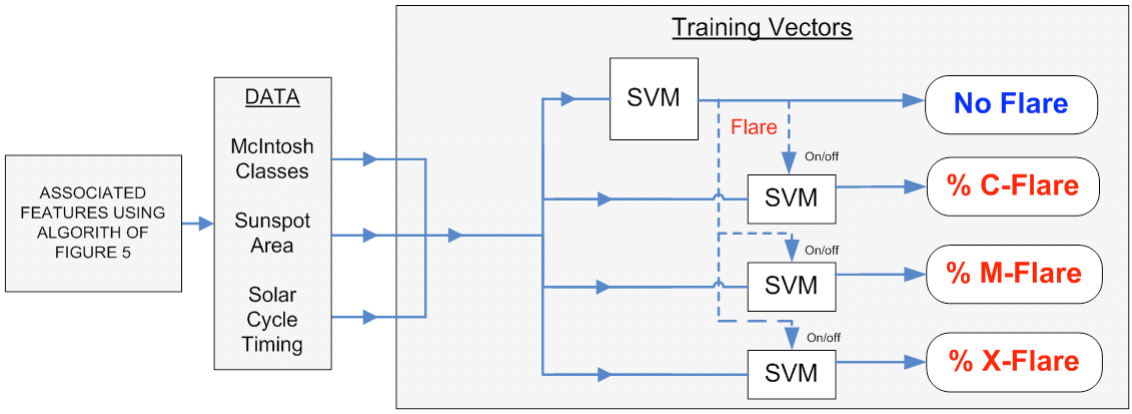


Figure 5.23 Flare Predictions - Learning Mode.

The best machine training performance was obtained when McIntosh Classes, Sunspot area, and normalized sunspot timing were used as inputs. After running the learning mode experiments 20 times, the resulting average prediction performance found is shown in Table 5-9.

The suggested improvements can be integrated with the Automated Solar Activity Prediction (ASAP) system proposed in Qahwaji and Colak (2006b), Colak and

Qahwaji (2007b), Colak and Qahwaji (2007c), and Qahwaji and Colak (2007) to provide a real-time prediction of solar flares as depicted in Figure 5.24.

Table 5-9 Flare Prediction - Learning Mode Results.

Prediction	Average Performance
Flaring	78.0%
C- Flare	79.5%
M-Flare	73.8%
X-Flare	80.8%

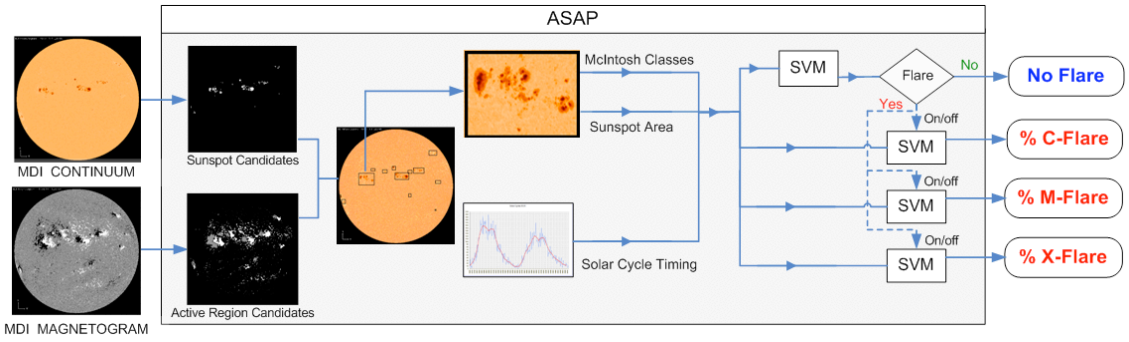


Figure 5.24 Suggested real-time mode for flare prediction.

Next, the characteristics of active regions that are associated with significant flares and CMEs were explored (Qahwaji et al., 2007b). The associations among these solar activities were found automatically as explained in Chapter 4. To apply machine learning algorithms, the input training vectors were created as shown in Figure 5.25. While optimizing the performance of the learning mode of the system for CME predictions, 9 inputs representing sunspots and flares features were used. The sunspot features are McIntosh Zurich, McIntosh Penumbra, McIntosh Distribution, Mt Wilson, normalized time, and the sunspot area, while the flare features are C/not-C, M/not-M, X/not-X.

As shown in Figure 5.25, the SVM system decides if there will be a CME or not (Qahwaji et al., 2007b). If a CME is predicted, then the CCNN machine is activated to predict its speed. The CCNN required 12 inputs, which consisted of the 9 inputs used for the SVM in addition to the normalized flare duration, the normalized flare decline

duration, and the normalised time difference between the flare peak time and the sunspot time.

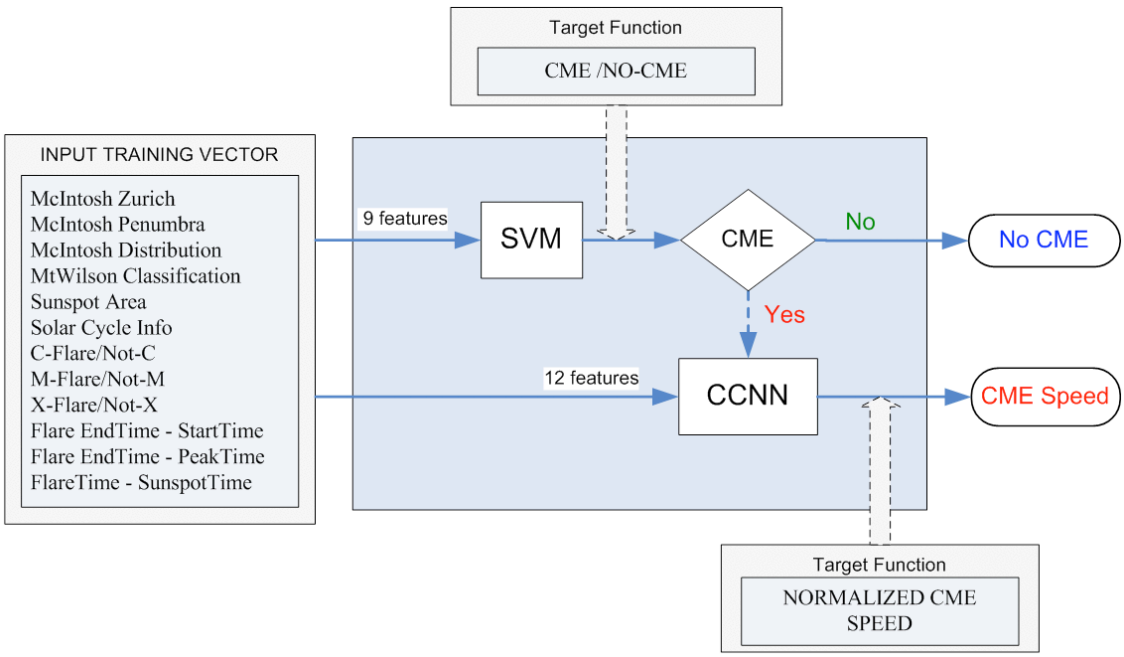


Figure 5.25 Learning mode of the CME prediction system (based on its associations with sunspots and flares).

Table 5-10 lists the results obtained by running 10 experiments for the system of Figure 5.25. The average CME prediction performance was found to be 64.4% and the average CME speed prediction performance was 73.9%.

Table 5-10 CME Prediction Learning Mode Results.

Experiment	SVM CME Prediction	CCNN SPEED Prediction
1	64.4%	73.3%
2	64.6%	71.4%
3	63.3%	78.0%
4	63.1%	75.6%
5	64.3%	71.9%
6	67.1%	72.4%
7	64.5%	72.9%
8	64.8%	74.3%
9	63.1%	74.8%
10	64.4%	74.6%
Average	64.4%	73.9%

The findings in this section show that real-time CME predictions can be provided by the integration between the computer platforms introduced in this chapter and the available flare prediction systems.

5.7 Performance Evaluation Comparisons

In the system that predicts CMEs based on their associations with flares (Section 0), the results show that SVM and CCNN perform better when three inputs are used: the intensity of the flare, flare duration and flare decline duration (Qahwaji et al., 2008c). It was found that a CCNN with 3 input nodes and 3 hidden nodes with a classification threshold of 0.56 gives the best results providing 0.63 *TPR* and 0.43 *FPR*. Also, a SVM classifier that accepts three inputs with d and γ values of 8 and 90 respectively and a classification threshold value of 0.83 provides the best prediction performance. This SVM configuration provides *TPR* and *FPR* values of 0.73 and 0.53 respectively. Comparing the results of SVMs and CCNNs it can be seen that the SVM classifier generates higher *TPR* compared to CCNN, but it also produces higher *FPR*. So, for a real-time system, choosing the right classifier will depend mainly on the objectives and domain of application of the system.

While investigating the associations between CMEs and filaments, it was found in the SVM experiments using validation method 1 that the best prediction performance had been achieved for input group 3, including filament timing, duration and type (Al-Omari et al., 2009a). From Table 5-6, this SVM configuration provides:

- Average *TPR* and *FPR* values of 0.65 and 0.22, respectively, which are seen from inspection of Figure 5.26 to provide better CME prediction performance than that obtained in the other experiments: using SVM in Subsection 5.5.2.1 (Al-Omari et al., 2008), using RBFs in Subsection 5.5.2.2 (Qahwaji et al., 2008a), using the Real and Modest AdaBoost in Subsection 5.5.3.1 (Qahwaji et al., 2008b), and using CCNN in

Subsection 5.4.2 (Qahwaji et al., 2008c). It is clear from the ROC curve of Figure 5.26 that the best prediction performance using SVM and flares associations in Subsection 5.4.3 (Qahwaji et al., 2008c) has a better *TPR* value than the use of SVM and filaments associations as it provided *TPR* value of 0.73, but with a high *FPR* value of 0.53. On the other hand, a more conservative performance was provided by the Gentle AdaBoost classifier presented in Subsection 5.5.3.1 (Qahwaji et al., 2008b), with *TPR* and *FPR* values of 0.46 and 0.12 respectively. Gentle AdaBoost is best used as a rejection classifier as it makes fewer false alarms.

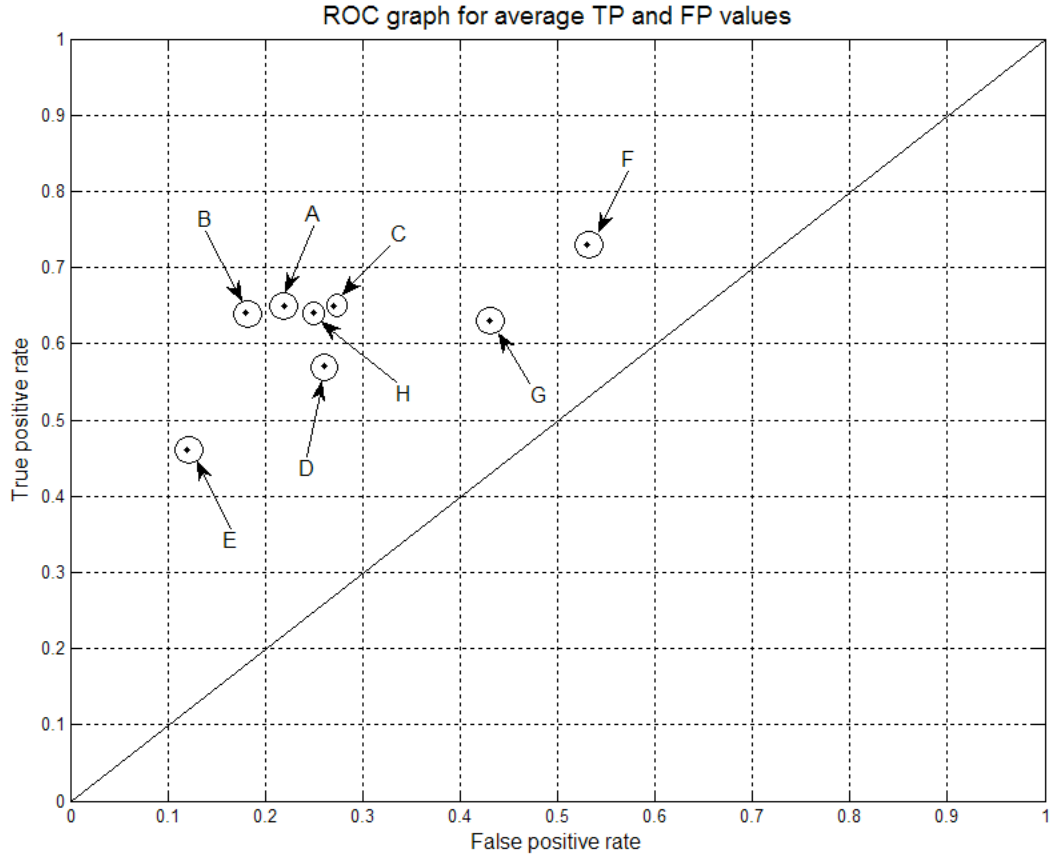


Figure 5.26 Comparisons among the prediction performances of the current work and all our previous research on CME prediction. {A} SVM-method 1 in Al-Omari et al. (2009a) {B} SVM-method 2 in Al-Omari et al. (2009a) {C} RBF in Qahwaji et al. (2008a) {D} Real and Modest AdaBoost in Qahwaji et al. (2008b) {E} Gentle AdaBoost in Qahwaji et al., (2008b) {F} SVM in Qahwaji et al. (2008c) {G} CCNN in Qahwaji et al. (2008c) {H} SVM in Al-Omari et al. (2008).

- An average accuracy of 73% which is the highest accuracy achieved so far in our research on predicting CMEs with the aid of the Jack-knife technique.
- An average *HSS* of 0.43, which is significantly better than random guessing. This value indicates that the system has forecasting ability and justifies confidence that the system is not predicting by chance or because of the statistical distribution of the selected data sample.
- A specificity (or *TNR*) of 78% which means a useful prediction performance if used as a rejection classifier to predict when CMEs are not likely to occur. A specificity of 88% has been achieved using the Gentle AdaBoost in Qahwaji et al. (2008b) but with a low *TPR* of 0.46. So, with an accuracy of 73% and specificity of 78% it is seen that our current system will be efficient if used as either a positive or a negative classifier tool for the purpose of CMEs prediction.

5.8 Conclusions

The findings in this chapter show that an increase in the CME prediction rate has been achieved with the use of more discriminative input features such as the flare decline duration and the filament type. The findings related to the insignificance of the flare incline duration can be explained by saying that the time needed for a flare to reach its peak intensity is not very important in terms of CMEs predictions using machine learning. It can be also said that the flare duration can be calculated as the sum of the incline and decline durations which means that the decline duration of the flare is very important for determining the probability of CME occurrence and this coincides with the findings of Yashiro et al. (2006).

A physical explanation for the strong relationship between the filament types and CMEs can be suggested from the Menzel-Evans classification (Menzel and Evans,

1953) where a filament/prominence is classified based on its material motion (upward or downward), its association with sunspots and its shape. From Figure 5.9 it is found that filaments with DSF, EPL and BSL types accounted for about 53.8% of the CME-associated filaments and these types are defined to be ascending from the Sun in their initial phase (Menzel and Jones, 1962). In addition, types like ASR (which rise above the limb) and BSD (which emanate from the Chromosphere) accounted for 12.9% of the CME-associated filaments. Hence, we conclude that filaments/prominences that originate from below in the Chromosphere (moving outward) are most likely to be associated with CMEs. On the other hand, it is reported that a loop prominence system (LPS) may appear as a flare in its initial phases (Jones, 1958) and the material in LPS prominences typically originates near the top of the loop and flows downward to the Sun. The association algorithm managed to associate only 2 LPS prominences with CMEs which suggests that filaments originating in the coronal space (moving downward) are not likely to be associated with CMEs.

All types of filaments/prominences occurring during solar cycle 18 (started in 1944 and ended in 1954) were investigated by Menzel and Jones (1962) who found that filaments/prominences originating in the coronal space (moving downward) represented 93.1% of the recorded prominences. This explains the low associations between CMEs and filaments in the findings of Chapter 4 and supports the conclusion that the direction of the material motion (upward or downward) of filaments can be used as an indicator for its association with CMEs.

It is shown in Figure 5.26 that the SVM experiments with validation method 2 (Subsection 5.5.3.3) has better performance compared to the first method using the Jack-knife technique (Subsection 5.5.3.2). From the results of both validation methods, it is clear that the CME prediction performance has been improved compared to the results of the other experiments. Checking some of the association cases manually

(using H-alpha images) and considering the mass loading model for CME initiation (conditions related to the speed and acceleration distributions of CMEs) enabled the association sets to be refined. Hence, it enabled the elimination of some cases of false associations, which produced some improvement in the prediction performance.

Overall, it is believed that the computerized learning rules extracted in this chapter can become an integrated part of many experimental studies in the field of forecasting, detecting, and classifying solar features and events. To achieve automated real-time CME predictions, it is intended to link this work with automated space weather systems that can provide real-time data for filaments and flares. For example, Bernasconi et al. (2005) introduced a computer system which provides automated detection, classification, and tracking for filaments in full-disk H-alpha images. Also, Scholl and Habbal (2008) proposed an automatic detection and classification system for coronal holes and filaments. On other hand, real-time data of solar flares can be extracted from the latest GOES X-ray flux profiles[12] or it can be obtained from the predictions provided by ASAP (Colak and Qahwaji, 2009).

In the next chapter, a statistical machine learning method will be used to study the evolution patterns of sunspot groups. Evolution patterns will be represented by computerised learning models to enable the next day prediction of the sunspot area and McIntosh classification. The complete system design for the automated real-time prediction system will be discussed in Chapter 7.

[12] http://www.swpc.noaa.gov/rt_plots, last access: 2009.

CHAPTER SIX

6 STUDYING THE SUNSPOT EVOLUTION PATTERNS USING HIDDEN MARKOV MODELS (HMMs)

6.1 Introduction

The main idea behind the work in this chapter (Al-Omari et al., 2009b, Qahwaji et al., 2009) is to model the evolution patterns of sunspot area and McIntosh classification using Hidden Markov Models (HMMs) in the form of computerised models (learning rules). These models can then be used to evaluate the likelihood of a given evolution sequence for the purpose of predicting its next-day state. Because of the close association between solar flares and Coronal Mass Ejections (Qahwaji et al., 2008c), and the strong relationship between sunspot regions and solar flares (McIntosh, 1990, Sakurai, 1970, Severny, 1965, Warwick, 1966), it has been decided to study the evolution of sunspot groups and incorporate this information in order to improve forecasts of flaring activity. The aims of this work are to:

- Investigate if the evolution patterns of sunspots can be modelled using computer-based HMMs.
- Investigate if the associations between sunspots and flares can be modelled using HMM such that flares can be predicted using a generalized model.
- Develop a model that can be used to predict the McIntosh class and the sunspot area for the sunspot group under investigation for the next 24 hours.

- Provide a future work plan on how the outcomes of this work can be used as part of more comprehensive work for the automated, near real-time prediction of flaring activity and sunspots evolution patterns for the next 24 hours.

The sunspot data used in this study were provided by Christopher Balch who is the lead forecaster of the space weather forecast office at the Space Weather Prediction Center (SWPC)[13]. As described in Chapter 3, the SWPC sunspot catalogue holds records including dates, locations, area, extent, McIntosh class, active region numbers (NOAA), and the class of associated solar flare events. The SWPC sunspot catalogue has been found to be very reliable and the data were consistent for HMM use because the sunspot records were equally spaced in time (one reading every 24 hours).

This chapter is organised as follows: Section 2 explores briefly HMMs and their properties. The Baum-Welch algorithm is summarised in Section 3. Section 4 describes the creation of the training and testing datasets and introduces the computer platform for a prediction system. The practical implementation and evaluation of the system is discussed in Section 5. Section 6 extends the work of this chapter in an attempt to model the associations between flares and sunspots. Finally, brief conclusions are provided in Section 7.

6.2 Hidden Markov Models (HMMs)

HMMs are used in this work for the following reasons:

- HMMs provide a mathematically consistent description of states and observations.
- HMMs are applicable to time-series predictions, which is the case when studying the evolution of sunspot patterns.

[13] http://www.swpc.noaa.gov/, last access: 2009.

- HMMs can be used as a classifier and at the same time provide modelling of the data.
- Models provided by HMMs can be easily validated.
- Learning by HMMs is still possible for data of variable-length vectors (number of features for data samples are not the same). That is while some sunspots can be observed during 13 consecutive days; other sunspots disappear within two days.

Rabiner (1989) provided a detailed tutorial on the use of HMMs in Speech Recognition. The history, theory, applications, and types of HMMs are explained very clearly in Rabiner (1989), Bengio (1999) and Kohlschein (2006).

6.2.1 HMM Parameters

To better understand the parameters of HMMs, assume that we have a hidden Markov model with a set of hidden states $S = \{S_1, S_2\}$ and a set of observations $O = \{O_1, O_2, O_3\}$. This model can be depicted as shown in the state diagram of Figure 6.1.

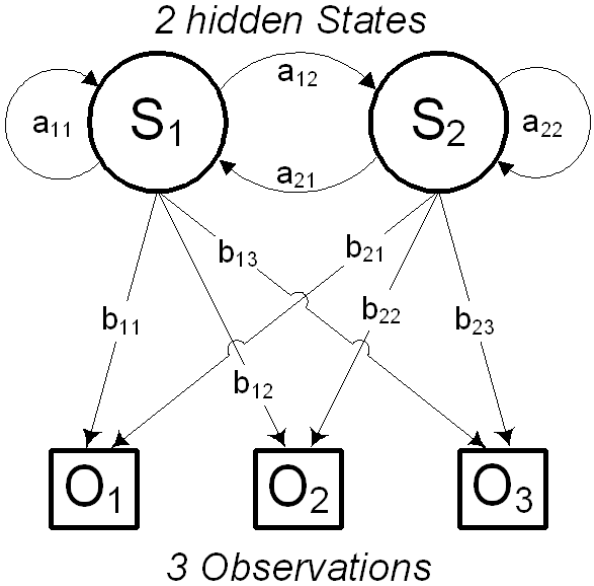


Figure 6.1 State diagram of a hidden Markov model showing its probabilistic parameters.

The model λ, shown in Figure 6.1, can be represented by $\lambda = (A, B, \pi)$ where A, B and π represent the following parameters:

- *Matrix of transition probabilities:*

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad (6-1)$$

$$a_{mn} = P(S_n \mid S_m) \quad ; \; m = 1,2 \quad n = 1,2 \qquad (6-2)$$

where a_{mn} is the probability that the current state is S_n given that the previous state is S_m. This is calculated as the expected number of transitions from state S_m to state S_n divided by the expected number of transitions out of state S_m.

- *Matrix of emission probabilities:*

$$B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} \qquad (6-3)$$

$$b_{np} \equiv b_n(p) = P(O_p \mid S_n) \quad ; \; n = 1,2 \quad p = 1,2,3 \qquad (6-4)$$

where $b_n(p)$ is the probability that the current observation is O_p given that the current state is S_n. It can be calculated as the expected number of times where O_p observed with S_n divided by the expected number of times in state S_n.

- *Initial states probabilities:*

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} \qquad (6-5)$$

$$\pi_m = P(S_m) \quad ; \; m = 1,2 \qquad (6-6)$$

where π_m is the expected number of times being in state S_m at the start time.

6.2.2 Challenges Associated with HMMs

Three main problems are associated with the use of HMMs:

1. *The evaluation problem*: Calculating the probability that a model $\lambda = (A, B, \pi)$ created a given sequence of observations.

2. *The decoding problem*: Finding the most likely sequence of hidden states, in a given model $\lambda = (A, B, \pi)$, that is created a given sequence of observations.
3. *The learning problem*: Estimating the model parameters $\lambda = (A, B, \pi)$ so that they best fit a given training sequences of observations.

More details and solutions for these problems are discussed in Rabiner (1989). The problem that is related to the work of this chapter is the learning problem (or parameter estimation). As a solution to this problem, the Baum-Welch algorithm is used for the reasons described in the following section.

6.3 Solution to the Learning Problem: The Baum-Welch Algorithm

In the learning problem, it is intended to maximize the probability of the training set (observations) given the model λ. So, the way the model parameters are optimised differs from one application to another based on the level of optimisation needed for that specific application. Generally, given any finite learning set, there is no optimal criterion for estimating the model parameters analytically (Rabiner, 1989). However, the model parameters can be adjusted such that the probability of the training observations is locally maximised for each class.

There are two main optimisation solutions reported in the literature for the learning problem: Maximum Mutual Information (MMI) and Maximum Likelihood (ML). Based on gradient techniques, the MMI consistently outperformed ML on speech recognition applications (Guo and Chan, 2006) because it is working to maximise the probability of the "word" string given the model parameters rather than the probability of the training observation sequences which can be referred to as Conditional Maximum Likelihood Estimation (CMLE) (Jurafsky and Martin, 2008). The ML criterion can be applied using iterative procedures, such as the Baum-Welch algorithm, or using gradient methods. Although gradient methods have proven to provide effective solutions for

similar applications (Rabiner, 1989, S.E.Levinson et al., 1983), the Baum-Welch algorithm has been shown to maximise the likelihood with a guaranteed convergence to a local maximum (Miklos and Meyer, 2005). Hence, it was decided to use the Baum-Welch algorithm in the current work.

As a solution to the learning problem, the Baum-Welch algorithm is used to adjust the model parameters $\lambda = (A, B, \pi)$ to best fit the observed data. If we have a training dataset of L observation sequences $V = V_1\ V_2\ \ldots\ V_L$ and a known values for the number of hidden states (N) and the number of possible observations (M), then we aim to maximize the term $P(V \mid \lambda)$.

The set of hidden states is $S = \{S_1, S_2, \ldots, S_N\}$ with the sequence $Q = q_1\ q_2\ \ldots\ q_t$ representing a sequence of hidden states up to time t. In addition, an observed sequence from the set of possible observations $\{O_1, O_2, \ldots, O_M\}$ can be represented by $O = o_1\ o_2\ \ldots\ o_T$ which is a sequence of T observations.

According to Rabiner (1989) and Gellert and Vintan (2006), the following variables need to be defined:

- $$\alpha_t(m) = P(o_1 o_1 .. o_t, q_t = S_m \mid \lambda) \qquad (6-7)$$

  which is the joint probability of the partial observation sequence up to time t and that the hidden state at time t is S_m given λ.

- $$\beta_t(m) = P(o_{t+1} o_{t+1} .. o_T \mid q_t = S_m, \lambda) \qquad (6-8)$$

  which is the probability of the partial observation sequence from time t+1 till T given λ and that the hidden state at time t is S_m.

- $$\xi_t(m,n) = P(q_t = S_m, q_{t+1} = S_n \mid o_1 o_2 ... o_T, \lambda) \qquad (6-9)$$

$$= \frac{P(q_t = S_m, q_{t+1} = S_n, o_1 o_2 ... o_T)}{P(o_1 o_2 ... o_T \mid \lambda)}$$

$$= \frac{\alpha_t(m) a_{mn} b_n(O_{t+1}) \beta_{t+1}(n)}{\sum_{m=1}^{N} \sum_{n=1}^{N} \alpha_t(m) a_{mn} b_n(O_{t+1}) \beta_{t+1}(n)}$$

which is the probability that the hidden state at time t is S_m and at time t+1 is S_n given the observation sequence and λ.

- $$\gamma_t(m) = P(q_t = S_m \mid o_1 o_1 ... o_T, \lambda) \qquad (6-10)$$

$$= \frac{P(q_t = S_m, o_1 o_2 ... o_T)}{P(o_1 o_2 ... o_T \mid \lambda)}$$

$$= \frac{\alpha_t(m)\beta_t(m)}{\sum_{m=1}^{N} \alpha_t(m)\beta_t(m)}$$

$$= \sum_{n=1}^{N} \xi_t(m,n)$$

which is the probability that the hidden state at time t is S_m given the observation sequence and λ.

As explained in Gellert and Vintan (2006), the Baum-Welch algorithm and the iterative Expectation Maximization (EM) algorithm are identical (have the same solution) for the current problem. Hence, the adjustment process for the parameters $\lambda = (A, B, \pi)$ is started as follows:

1. Initialize the parameters $\lambda = (A, B, \pi)$ randomly: a_{mn} is initialized to $1/N$, b_{mp} is initialized to $1/M$, and π_m is initialized to $1/N$.
2. From the equations (6-7) through (6-10), calculate the parameters $\alpha_t(m)$, $\beta_t(m)$, $\xi_t(m,n)$ and $\gamma_t(m)$.
3. Calculate the new parameters of the model $\lambda^* = (A^*, B^*, \pi^*)$ according to the values calculated in step 2 as follows:

$$a^*_{mn} = \frac{\sum_{t=1}^{T} \xi_t(m,n)}{\sum_{t=1}^{T} \gamma_t(m)} \qquad (6-11)$$

$$b^*_n(p) = \frac{\sum_{\substack{t=1 \\ o_t = O_p}}^{T} \gamma_t(n)}{\sum_{t=1}^{T} \gamma_t(n)} \qquad (6-12)$$

$$\pi^*_m = \gamma_1(m) \qquad (6-13)$$

4. Calculate $P(V \mid \lambda^*)$. While the probability $P(V \mid \lambda^*)$ is increasing repeat steps 2 and 3.

After the model parameters converge to some values, these parameters will be describing a model that best fits the training observation sequences.

6.4 The Prediction System Design

This chapter introduces a computer platform that uses HMMs for studying the evolution patterns of sunspot areas and McIntosh classifications. As described in Al-Omari et al. (2009b), the prediction models are validated as shown in Figure 6.2.

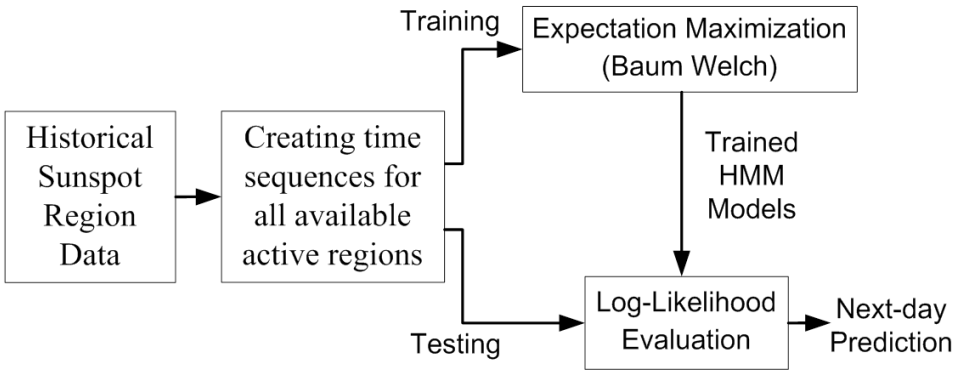


Figure 6.2 The validation process for our next-day sunspot area and McIntosh classification prediction system.

6.4.1 Tracking Active Region Data

The system in Figure 6.2 starts by tracking the sunspot area and McIntosh classifications for all active regions in the period between 18/08/1996 and 31/03/2006 using their NOAA numbers. A total of 2876 active regions have been tracked by their NOAA numbers and two datasets have been created. The first dataset contains records for the daily values of sunspot areas as shown in Table 6-1 and the second dataset contains the corresponding McIntosh classifications for these sunspots as shown in Table 6-2.

Table 6-1 Evolution dataset for sunspot areas (in millionths of solar hemisphere).

NOAA	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14
7986	50	90	100	230	60	50	100	60	50	30	30	20	0	
8084	10	10	10	70	170	270	240	240	230	270	240	270	80	
10486	150	1160	1540	2200	2170	2180	2120	2610	2600	2030	1900	2160	1430	630
10487	160	250	240	170	150	280	130	110	80	40	20	20		

Table 6-2 Evolution dataset for McIntosh classifications.

NOAA	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14
7986	AXX	HSX	CSO	CSO	HSX	HSX	CSO	CSO	CSO	HSX	HSX	HRX	AXX	
8084	BXO	HRX	HRX	CAO	DAI	DSI	EAI	EAO	EAO	ESO	DAO	EAO	DRO	
10486	HKX	EKC	FKC	FKC	FKC	FKC	FKC	FKC	FKC	FKC	FKC	FKC	FKC	EKC
10487	CSO	DAO	DAO	DAO	DAO	DKO	DAI	DAO	DAO	CSO	CSO	AXX		

Data of active regions that last for three days or less were excluded from the datasets. Hence, the final datasets consisted of data representing 2101 active regions. Then, these datasets were processed to find all observation sequences that consist of three days data (or more) plus the next day data. For example, the available data for active region 10487, shown in Table 6-2, were used to create 9 observation sequences as shown in Table 6-3. Finally, a total of 12693 observation sequences were created for the McIntosh classification data with another 12693 sequences corresponding to the sunspot areas.

Table 6-3 Observation sequences extracted from the McIntosh classification dataset for active region 10487.

Observation Sequence	Next Day Class
CSO, DAO, DAO	DAO
CSO, DAO, DAO, DAO	DAO
CSO, DAO, DAO, DAO, DAO	DKO
CSO, DAO, DAO, DAO, DAO, DKO	DAI
CSO, DAO, DAO, DAO, DAO, DKO, DAI	DAO
CSO, DAO, DAO, DAO, DAO, DKO, DAI, DAO	DAO
CSO, DAO, DAO, DAO, DAO, DKO, DAI, DAO, DAO	CSO
CSO, DAO, DAO, DAO, DAO, DKO, DAI, DAO, DAO, CSO	CSO
CSO, DAO, DAO, DAO, DAO, DKO, DAI, DAO, DAO, CSO, CSO	AXX

6.4.2 Creating the Numerical Sequences

The next step in the validation process is to train the data using the Baum-Welch algorithm as described in the previous section. But before that we need to represent the

observation sequences numerically. Then the Baum-Welch algorithm can be used to train many HMMs, one for each possible area level and one for each possible McIntosh classification.

As described in Chapter 3, the McIntosh classification consists of three components: the sunspot class (modified Zurich class), the penumbral class (type of largest spot), and the sunspot distribution (degree of compactness in the interior of the group). According to McIntosh (1990), there are 60 allowed combinations for the individual components of the McIntosh classification system. Hence, two options are available for the numerical representation of the McIntosh classes:

- To use integers from 1 to 60.
- To extract the individual components of the classification and represent them separately. So, the sunspot classes A, B, C, D, E, F, and H can be represented by the integers from 1 to 7, respectively. In the same manner, the penumbral classes X, R, S, A, H, and K can be represented by the integers from 1 to 6, respectively. And the sunspot distributions X, O, I, and C can be represented by 1, 2, 3, and 4, respectively.

Observation sequences of the McIntosh classification were represented using both options and practical experiments were carried out to determine the best option, as explained later. Sunspot areas were quantized in 10 levels. Each level corresponds to 100 millionths of the solar hemisphere (S_H). The quantized levels are centred at {50, 150, 250, ..., 950} millionths of S_H such that level 1 represents sunspot areas from 0 to 100 millionths S_H, level 2 represents sunspot areas from 100 to 200 millionths S_H and so on. All regions that have areas greater than 900 millionths S_H are assigned to level 10.

6.5 Practical Implementation and Results

6.5.1 Training-Testing Experiments

To validate the system, 10154 observation sequences (80% of the available data) were selected randomly for the training of HMM models. The rest of the data (2539 sequences) were used for testing. The training was processed using the Baum-Welch algorithm to:

1. Train 60 HMMs, one for each McIntosh classification.
2. Train seven HMMs, one for each sunspot class.
3. Train another six HMMs, one for each penumbral class.
4. Train four more HMMs, one for each sunspot distribution.
5. Train 10 HMMs, one for each area level.

The testing process was done as follows: if the sunspot data is given for at least three previous days, then the likelihood of its evolution sequence is evaluated for all the trained HMMs. Then predictions are given based on the sunspot area level and the McIntosh classification that provides the maximum likelihood. The HMM Toolbox for MATLAB, written by Kevin Murphy[14], was used in these experiments.

Initially, the training-testing experiments were carried out for the McIntosh classification using 60 observations. The number of hidden states was increased from 5 to 100 in steps of 5 and for each hidden state value the training-testing experiments were repeated five times with different sets of observation sequences. The averages of the correct predictions obtained are shown in Table 6-4.

It is found from Table 6-4 that HMMs cannot be usefully trained to best fit the observation data while representing the McIntosh classes using 60 states. This is clear from the result that the best prediction performance achieved is no more than 36.5%. Hence, instead of training HMMs with a set of 60 possible observations, it was decided

[14] http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html, last. access: 2009.

to analyse the individual McIntosh components and train them separately. This means that 17 HMMs (seven sunspot classes + six penumbral classes + four sunspot distributions) for the McIntosh classes and 10 HMMs for the possible area levels are needed for training. The complete training process is depicted in Figure 6.3.

Table 6-4 Results for the McIntosh class prediction using 60 observation states.

Number of Hidden States	Average Correct McIntosh Predictions [%]
5	32.0
10	34.9
15	34.6
20	30.9
25	35.3
30	35.8
35	34.0
40	33.9
45	35.9
50	33.9
55	34.6
60	36.5
65	35.2
70	35.2
75	34.7
80	35.8
85	34.5
90	33.7
95	33.1
100	34.8

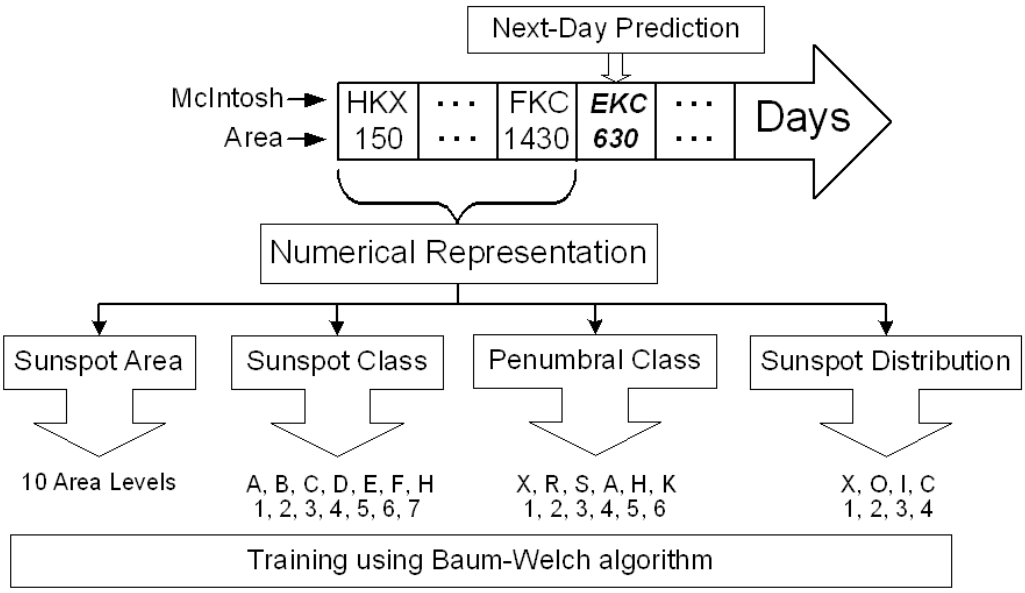


Figure 6.3 The training mode of the prediction system.

6.5.2 Validation Results

At the end of the training experiments, 27 trained HMMs were obtained representing all defined area levels and McIntosh classes. Then the likelihood of each sample in the testing data was evaluated for each one of these models. The training-testing experiments were repeated 10 times and each time the training and testing datasets were selected randomly. With 60 hidden states, the average accuracy of the next-day prediction of area and McIntosh classification was found to be around 71% and 60% respectively as shown in Table 6-5.

Table 6-5 Results for 10 experiments with 60 hidden states.

Exp.	Correct McIntosh Predictions [%]	Correct Area Predictions [%]
1	59.0	71.4
2	60.9	72.0
3	59.5	71.2
4	59.4	68.4
5	58.3	70.7
6	58.9	68.3
7	59.4	72.7
8	59.7	69.6
9	60.2	69.8
10	59.8	71.1
Average	59.5	70.5

The same training-testing procedure was repeated while changing the number of hidden states from 5 to 60 in steps of 5 and it was found that the best results were obtained with 60 hidden states as shown in Table 6-6. However, it can be concluded that changing the number of hidden states does not have a significant effect on the prediction results which suggests the use of the most parsimonious model with 5 hidden states. It is worth mentioning here that the average time needed to train 17 HMMs, representing McIntosh components, was about five minutes with 5 hidden states while it took more than nine minutes to train the same models using 60 hidden states. In contrast, the training experiments in the previous subsection (Table 6-4) took about four minutes to train 60 HMMs with 10 iterations each.

Table 6-6 Average results with different numbers of hidden states.

Number of Hidden States	Average Correct McIntosh Predictions [%]	Average Correct Area Predictions [%]
5	59.4	70.2
10	59.3	68.8
15	59.2	69.2
20	59.1	67.9
25	59.3	68.4
30	59.5	67.9
35	59.2	69.4
40	59.4	68.8
45	59.1	68.9
50	59.4	70.1
55	59.1	69.7
60	59.5	70.5

It was found that representing the McIntosh classes using 17 HMMs (the individual components) provided better results than the representation using 60 HMMs. In addition, the testing experiments while considering the individual components separately were found to be faster than testing the 60 possible McIntosh classes. This is because a lower number of likelihood calculations is needed for 17 HMMs.

On the other hand, the training process for 60 HMMs was found to be faster than that for 17 HMMs. An explanation for this can be drawn from the frequency of McIntosh classes shown in Table 6-7 and Table 6-8. Generally, it can be noted that there is a lower number of training sequences for each HMM of the 60 models compared to the number of training sequences for each of the 17 HMMs.

Table 6-7 Frequency of McIntosh classifications over the data used in the training-testing experiments.

McIntosh	Sunspot Class							Penumbral Class						Sunspot Distribution			
	A	B	C	D	E	F	H	X	R	S	A	H	K	X	O	I	C
Frequency	1489	1899	4004	4664	2125	844	3971	3388	1035	6781	5710	301	1781	5460	10933	1892	711

Table 6-8 Frequency of McIntosh classifications over the data used in the training-testing experiments.

McIntosh	Frequency	McIntosh	Frequency	McIntosh	Frequency
AXX	1489	DKI	163	ESO	250
BXI	3	DKO	181	FAC	31
BXO	1896	DRI	6	FAI	141
CAI	17	DRO	115	FAO	98
CAO	1102	DSC	4	FHC	10
CHI	2	DSI	74	FHI	10
CHO	68	DSO	1513	FHO	18
CKI	3	EAC	68	FKC	217
CKO	120	EAI	435	FKI	184
CRI	1	EAO	572	FKO	64
CRO	579	EHC	3	FRI	0
CSI	4	EHI	24	FRO	1
CSO	2108	EHO	26	FSC	3
DAC	61	EKC	223	FSI	18
DAI	435	EKI	309	FSO	49
DAO	1970	EKO	152	HAX	780
DHC	3	ERI	0	HHX	81
DHI	7	ERO	2	HKX	82
DHO	49	ESC	5	HRX	331
DKC	83	ESI	56	HSX	2697

6.6 Real-Time System

In Colak and Qahwaji (2007a), sunspot groups were detected and classified automatically by analysing their complexity and area using a system called “ASAP” (Colak and Qahwaji, 2009). ASAP provides automated real-time sunspot classifications and flare predictions as shown in Figure 6.4 (Colak and Qahwaji, 2007b, Colak and Qahwaji, 2007c, Colak and Qahwaji, 2007a). Based on ASAP’s classifications, the evolution of sunspot groups can be studied and predicted (Al-Omari et al., 2009b, Qahwaji et al., 2009).

ASAP is the first automated space weather prediction system, but it has the following limitations: it does not track sunspots over long periods of time, and it does not study the evolution pattern of the detected sunspots.

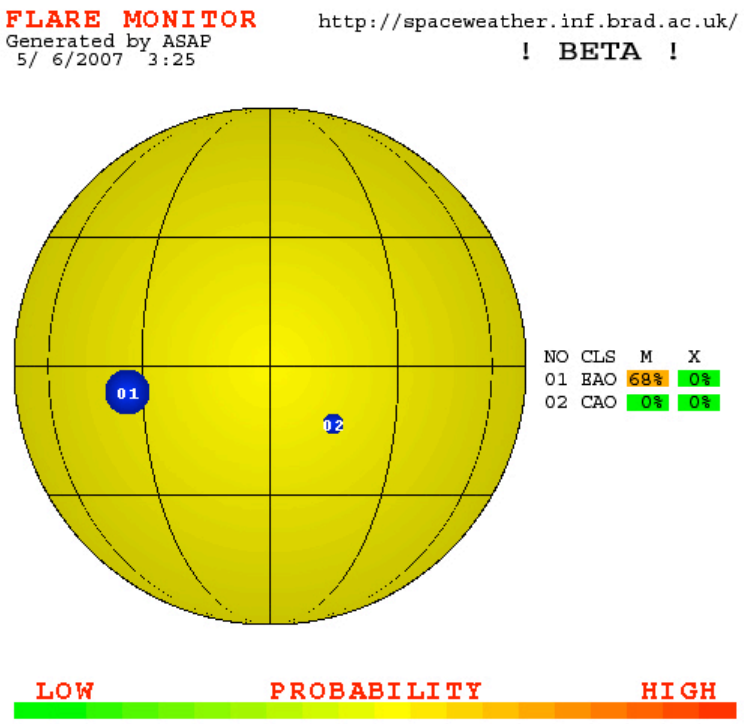


Figure 6.4 An example on ASAP's detections, classifications, and predictions.

The technology introduced in this chapter has the potential to overcome one of these limitations of ASAP because it is able to extract the evolution patterns for sunspot groups over days of observations. However, further work is needed to design the second generation ASAP to properly integrate this new technology with ASAP's existing systems and overcome compatibility problems and re-train the whole system. This system once completed will be similar to the system shown in Figure 6.5 and will be able to provide near-real time prediction for the evolution patterns of sunspots.

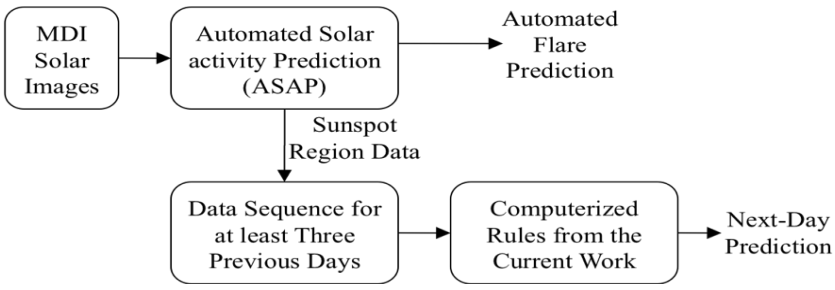


Figure 6.5 Near-real time prediction for the evolution patterns of sunspot regions.

A complete future work plan is provided in the next section including suggestions on how this work can be improved and extended for the purpose of real-time prediction of solar flares.

6.7 Modeling the Associations between Sunspot Groups and Flares

Because this work is the first to use HMMs in space weather forecast studies, it is intended to study possible extensions of the proposed design in future work. In Qahwaji et al. (2007b) the evolution of active region number 10486 has been studied manually as shown in Figure 6.6.

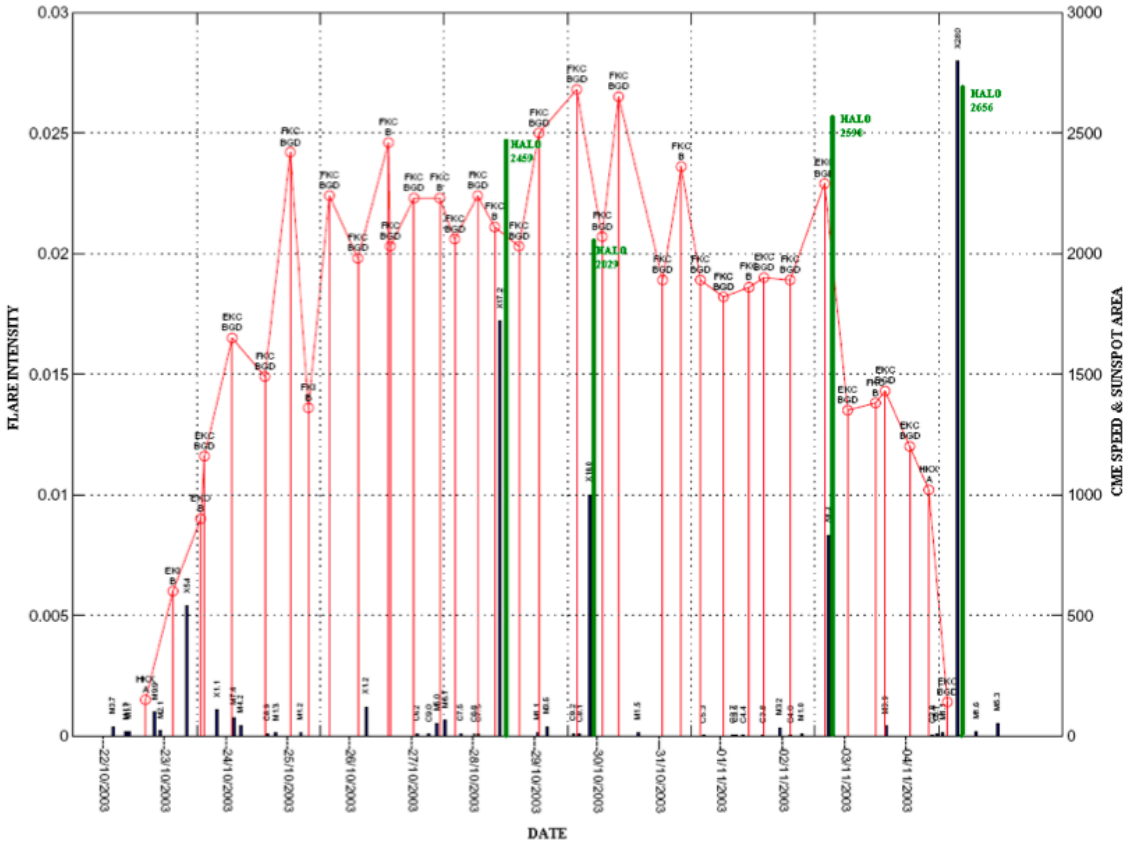


Figure 6.6 The evolution of AR10486 and its associated flares and CMEs.

This active region caused the largest solar X-ray flare recorded, which was classified as X28 and occurred on 14 November 2003. It was also one of the most active regions observed during the Halloween Storm, which occurred late October and early November 2003. From Figure 6.6, which shows the evolution of AR10486, it is clear that there is a strong association between the sunspot group area (red) and the McIntosh classification, solar flare intensity (black), and CME speed (green). From the general trend of the data in Figure 6.6 it can be noted that most of the fast CMEs are initiated after the peak time of significant X or M-class flares which are associated with sharp

decreases in the sunspot areas and McIntosh classifications of FKC or EKC. To draw a general conclusion from this study, it would be necessry to repeat the same graphical analysis for all the available active region data, which is impractical to carry out manually. It is believed that the HMM can be used to investigate whether there is an evolution pattern that is repeated for eruptive active regions.

As discussed previously, solar flare research has shown that flares are mostly related to sunspots and active regions (Liu et al., 2005, Shi and Wang, 1994, Zirin and Liggett, 1982). The HMM system described in Section 6.4 can be redesigned to search for a model that can provide a complete representation of the relation between a sunspot region evolution and its associated flaring activity. For example, tracking all the available data for active region 10486, along with its associated flaring activity, provides the sequences shown in Table 6-9.

Table 6-9 Evolution sequences for AR10486 and its associated flaring activity.

Days	Area	McIntosh	Flaring
1	150	HKX	M
2	1160	EKC	X
3	1540	FKC	M
4	2200	FKC	M
5	2170	FKC	X
6	2180	FKC	M
7	2120	FKC	X
8	2610	FKC	X
9	2600	FKC	M
10	2030	FKC	C
11	1900	FKC	M
12	2160	FKC	X
13	1430	FKC	M
14	630	EKC	X

For the purpose of modelling the association between sunspots and flares, the observable sunspots data are considered as observations and the classes of the associated flares are considered as the hidden states. The set of hidden states is $F = \{1, 2, 3, 4\}$ where the integers from 1 to 4 represents “No Flare”, “C Flare”, “M Flare” and “X Flare”, respectively. The matrix of transition probabilities A_F and initial states

probabilities π_F can represent the evolution of the flaring activity in an active region. Initial values of A_F and π_F can be calculated from the SWPC sunspots catalogue, in the period between 18/08/1996 and 31/03/2006, as follows:

$$A_F = \begin{bmatrix} 0.885 & 0.100 & 0.015 & 0.000 \\ 0.486 & 0.397 & 0.110 & 0.008 \\ 0.261 & 0.385 & 0.294 & 0.060 \\ 0.060 & 0.301 & 0.458 & 0.181 \end{bmatrix} \qquad (6-14)$$

$$\pi_F = \begin{bmatrix} 0.915 \\ 0.067 \\ 0.016 \\ 0.001 \end{bmatrix} \qquad (6-15)$$

The values in A_F and π_F can provide a general indication of the flaring activity evolution in active regions. For example, active regions tend to have no flaring activity (probability = 0.915) during their initial stages with a probability of 0.067 of producing C-class flares. If an active region is associated with an M-class flare, then the probability that it will produce an X-class flare the next day is 0.060.

To better represent the observable McIntosh classifications, three HMMs are suggested: HMM1 for sunspot class (Figure 6.7), HMM2 for penumbral class (Figure 6.8), and HMM3 for sunspot distribution (Figure 6.9).

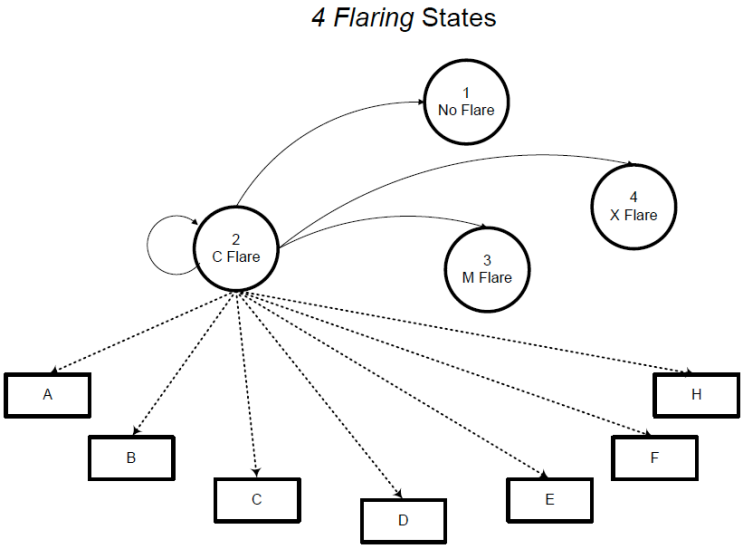


Figure 6.7 Sunspot Class state diagram (HMM1).

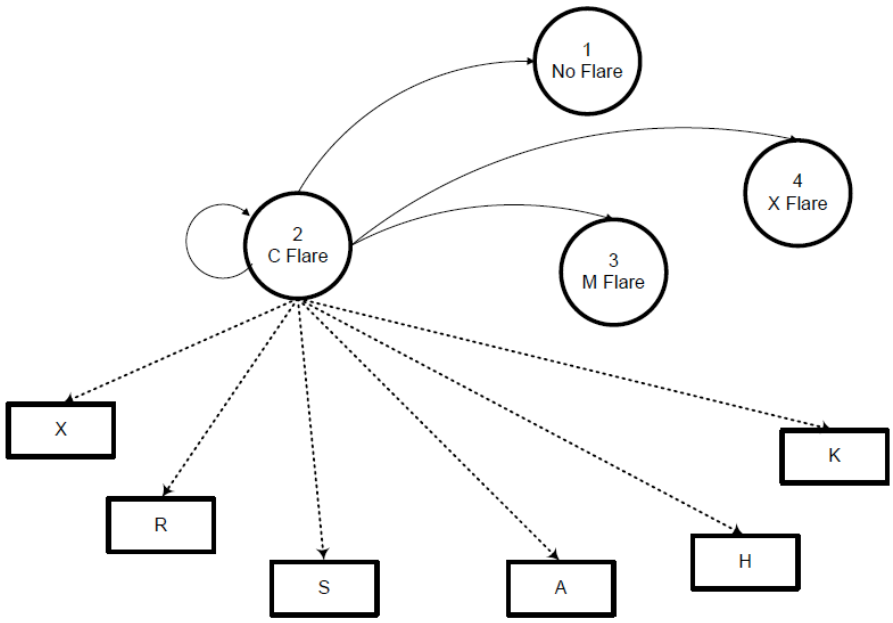


Figure 6.8 Penumbral Class state diagram (HMM2).

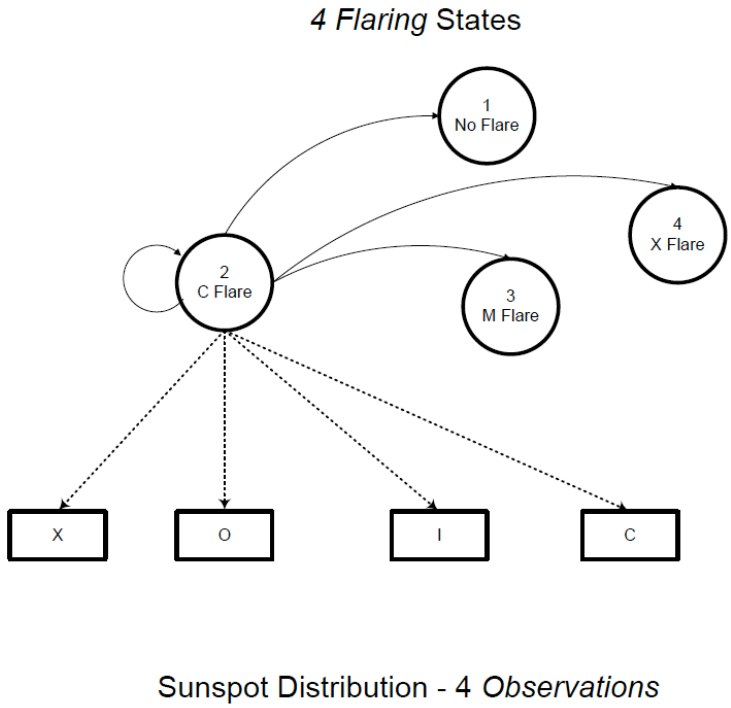


Figure 6.9 Sunspot Distribution state diagram (HMM3).

The set of observations for HMM1 are $O_1 = \{A, H, B, C, D, E, F\}$. The initial value of the matrix of emission probabilities B_1 is calculated from the SWPC sunspots catalogue as:

$$B_1 = \begin{bmatrix} 0.118 & 0.233 & 0.139 & 0.224 & 0.210 & 0.062 & 0.015 \\ 0.016 & 0.064 & 0.039 & 0.158 & 0.340 & 0.262 & 0.112 \\ 0.004 & 0.046 & 0.009 & 0.084 & 0.306 & 0.316 & 0.235 \\ 0.011 & 0.032 & 0.000 & 0.021 & 0.223 & 0.362 & 0.351 \end{bmatrix} \quad (6-16)$$

The set of observations for HMM2 are $O_2 = \{X, R, S, A, H, K\}$ and for HMM3 are $O_3 = \{X, O, I, C\}$. In the same way, the initial values of the matrices of emission probabilities B_2 and B_3 are calculated as:

$$B_2 = \begin{bmatrix} 0.257 & 0.066 & 0.374 & 0.250 & 0.011 & 0.042 \\ 0.055 & 0.029 & 0.234 & 0.429 & 0.030 & 0.224 \\ 0.013 & 0.015 & 0.136 & 0.410 & 0.030 & 0.395 \\ 0.011 & 0.000 & 0.032 & 0.160 & 0.011 & 0.787 \end{bmatrix} \quad (6-17)$$

$$B_3 = \begin{bmatrix} 0.351 & 0.593 & 0.047 & 0.009 \\ 0.079 & 0.547 & 0.264 & 0.109 \\ 0.050 & 0.400 & 0.335 & 0.216 \\ 0.043 & 0.181 & 0.255 & 0.521 \end{bmatrix} \quad (6-18)$$

To provide a clear view of the relationship between the individual McIntosh components and the flaring activity, the emission matrices B_1, B_2, and B_3 are depicted in Figure 6.10, Figure 6.11, and Figure 6.12.

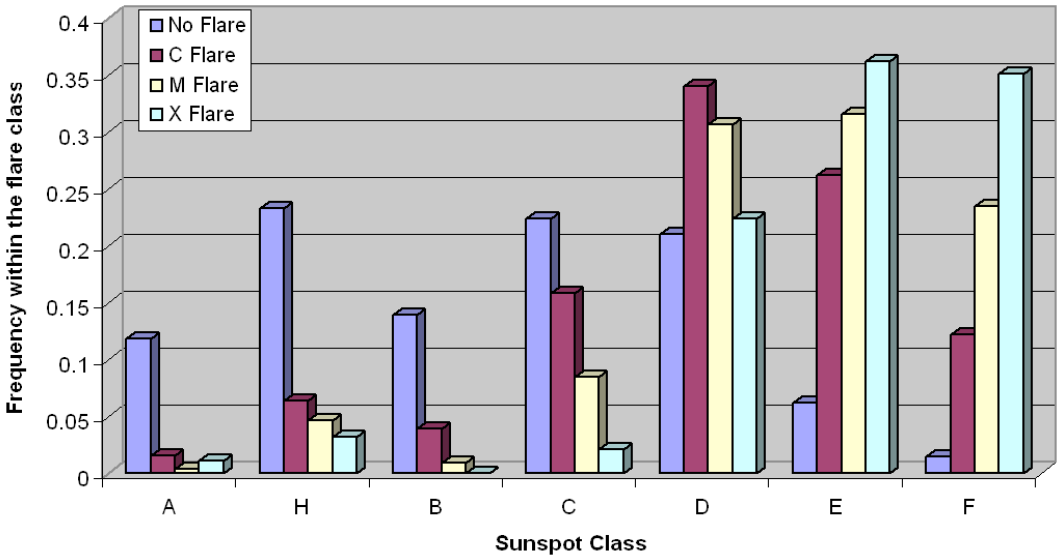


Figure 6.10 Sunspot classes vs flare classes.

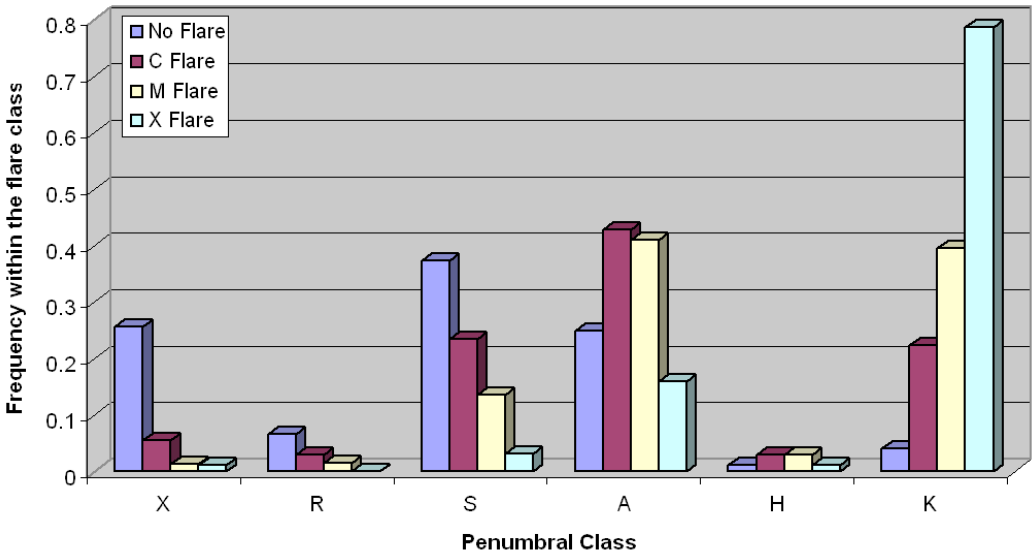


Figure 6.11 Penumbral classes vs flare classes.

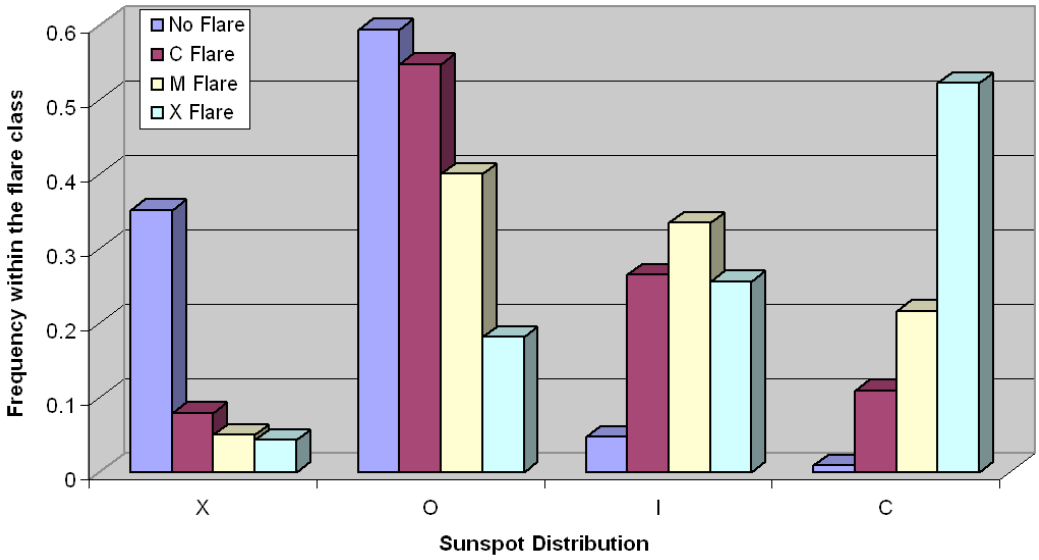


Figure 6.12 Sunspot distribution vs flare classes.

It is clear that there is a strong relationship between the flaring activity and the individual components of the McIntosh classification. For example, if the sunspot distribution component of McIntosh is classified as C, then it can be used as a strong indicator of the possibility of the creation of an X class flare. This possibility increases if the penumbral class is K (McIntosh class of FKC or EKC). It is believed that the suggested models can be used in the future to provide computerised HMMs and the outcomes of such work can be integrated with ASAP to provide an enhanced long-term (days) prediction of solar flares. In addition this integration could enable the real-time

CME predictions based on the associations between CMEs and flares as explained in the previous chapter.

6.8 Conclusions

The system design presented in this chapter provides computerized Markovian models that are applicable for sequential solar data. The sunspot catalogue provided by SWPC is found to be suitable for the use of HMMs because it has a fixed time period of 24 hours between each sunspot record and the next one (for each active region). However, it is concluded that one record a day is not enough to satisfactorily describe the evolution patterns of an active region.

It is shown by the validation experiments that it is better to predict the next-day McIntosh classification from the individual components. This corresponds to lower number of HMMs and faster evaluation for the likelihood values.

It is believed that the system design presented in this chapter is the first to provide a computerized large-scale study for the evolution patterns of sunspot regions. However, improvements are still needed. For the near future it is planned to continue this work and create tools that could identify the observed patterns in recent images and display similar patterns of evolution from past historical cases. This work can also be integrated with ASAP using data fusion techniques to enhance the automated flare predictions generated by ASAP.

CHAPTER SEVEN

7 CONCLUSIONS AND SUGGESTIONS FOR FURTHER WORK

7.1 Conclusions

7.1.1 Overall Conclusions

The main achievement of the research presented in this thesis can be described as the designs and associated validation discussions and conclusions of a collection of systems related to the field of space weather forecasting, which have been developed to provide machine learning based technologies and computerized decision rules and models. These rules and models have been trained and tested using historical solar data representing properties of different solar features and events including sunspot groups, filaments/prominences, solar flares, and CMEs. It is believed that this work is important because for the first time a machine learning-based, large-scale investigation for the associations between these solar features and events have been explored and verified for years of data. Also, for the first time, the evolution of sunspot groups is studied and fitted into a time-series model using Hidden Markov Models (HMMs). These associations and models have been represented using computerised learning rules. This representation is an important step for creating automated and reliable prediction systems that can predict the extremes of space weather.

7.1.2 Detailed Conclusions

Concluding remarks on this research are listed as follows:

- A computer tool described in Chapter 4 was developed to study the associations among CMEs, flares, filaments, and sunspots by processing the NGDC sunspot, filament and flare catalogues and the SOHO/LASCO CME catalogue. The fixed data formats of these catalogues facilitated their automatic processing by the proposed association tool. In addition, they are produced regularly, are easy to access and analyse and represent important information regarding solar features without the need to process the large-size solar images. On the other hand, it was concluded in Chapter 4 that there are many events reported in these catalogues with missing properties. For example, the flare locations and NOAA numbers were not reported in about half the investigated flares and many sunspot records were reported without McIntosh classifications or areas. Also the locations, extents, and NOAA numbers were missing for a large number of filament records. It was intended to use the SWPC sunspots catalogue in the study of associations but unfortunately only one sunspot record a day per active region is reported in this catalogue which is not enough to represent the actual activity.
- The associations between CMEs and solar flares, reported in the period between the years 1996 to 2004, were studied using the association tool. The highest association rate was found to be between CMEs and X-class flares where CME candidates were found for 68.3% of the X-class flares based on their timing information. On the other hand, only 38.8% of the M-class and 18.1% of the C-class flares were associated with CMEs. Based on these associations it was concluded that there is a direct relation between the initiation of a CME and the X-ray intensity for the associated flaring activity and it was found that the higher the X-ray intensity, the faster the CME would be initiated.

- The tool was used to find the associations between CMEs and filaments/prominences. It managed to associate 6.3% of the filaments with 6.1% of the CMEs, reported in the period between years 1996 to 2001, based on their timing and location information. It was concluded that there is a strong relation between CMEs and the material motion direction of filaments. Certain filament types such as DSF, EPL, BSL, ASR and BSD account for about 66.7% of the CME-associated filaments and these defined types were either emanating from the chromosphere or moving outward. These association findings were found to be supported by Moon et al. (2002) and the low association rates were found to be explainable by the findings of Menzel and Jones (1962) who found that filaments moving outward represented only 6.9% of the recorded events during solar cycle 18.
- In Chapter 5 different machine learning algorithms and different types of flare and filament properties were compared for CME prediction. The association datasets were processed in appropriate numerical format so that they could be processed by machine learning algorithms. It was concluded that the most important input features for a CME prediction system were the flare intensity and decline duration (the time duration from the flare's peak intensity to the end of the flare) and the filament type and duration. These features were shown to be good indicators for the possibility of initiating CMEs.
- For the first time, CCNNs and SVMs were compared within the context of CME predictions based on CME-flare associations. Flare intensity, duration and decline duration were used as inputs to determine if a flare is going to initiate a CME. After conducting extensive experiments, it was found that CCNNs provide a more conservative CME predictions performance with 0.63 *TPR*, 0.43 *FPR* and 60.4% accuracy compared to the more "liberal" performance (Fawcett,

2006) of 0.73 *TPR*, 0.53 *FPR* and 62.4% accuracy using SVMs. It was concluded that the SVM classifier provided better performance in terms of accuracy and correct positive predictions compared to CCNN, but it also produced higher rates of false alarm predictions.

- Several machine learning algorithms were optimised and compared to analyse the CME-filament associations and to provide CME predictions. The best CME prediction performance for the datasets considered were obtained using SVMs which were validated using the Jack-knife technique. It was concluded that if the real-time properties of an observed filament (solar cycle time, duration, and type) are available, then the SVM learning rules could be used to predict if this filament is going to initiate a CME with true positive prediction probability of 65%. At the same time, it could be predicted if there will be no CME initiated by the observed filament with a true negative prediction probability of up to 78%. So, the whole system, when used for predicting CMEs, can achieve a correct prediction probability of 73%.
- The work of Qahwaji and Colak (2007) was used as a starting point to study the relation between the CME-flare associated pairs and characteristics of the corresponding sunspot groups. These Authors considered M- and X-class flares covering years of data from both solar cycles 22 and 23. Their work was extended to include C-class flares covering most of solar cycle 23 and it was then integrated with the CME-flare association algorithm under a new system that can predict CMEs by analyzing their associations with sunspots and flares using machine learning. From the initial work on this design it was found that if real-time sunspot information (McIntosh, MtWilson and area) and flaring activity information (class and timing) are available, then CME predictions can be provided using SVMs with an accuracy of up to 64.4%. Also, it was

concluded that there is a strong associations between fast CMEs, initiated within the decline duration of significant X or M-class flares, and certain McIntosh classes such as FKC, EKC and FKI.

- Overall it was concluded that SVMs, which were originally designed for binary classifications, provided the most accurate and reliable prediction performance. In addition, SVM training-testing experiments were found to be the fastest compared to other learning algorithms. On the other hand, CCNNs needed most training time. These results support the comparison findings in Qahwaji and Colak (2007). It is worth mentioning that the Gentle AdaBoost provided the best prediction performance as a rejection classifier (predicting when CMEs are not likely to occur) with a specificity (or *TNR*) of 88%. It is believed that choosing the best classifier depends mainly on the objectives and domain of application.
- For the first time, the possible evolution patterns of sunspot groups were studied using HMMs (Chapter 6). An HMM-based technology was developed to model these evolution patterns along with the possible flaring activity to provide long-term predictions for sunspot areas, McIntosh classifications, and flaring activities. With the availability of the sunspot area and McIntosh classification for at least three previous days, it was found that the next-day area and McIntosh class can be predicted with accuracies of 71% and 60% respectively.
- In an attempt to model the associations between solar flares and sunspot groups using HMMs, initial work was presented in Chapter 6 (as a future work plan). It was proven statistically that the individual components of the McIntosh classification could be used separately as an indicator for the possible flaring activity. From the matrices of emission probabilities calculated in Section 6.7, it was found that the highest probabilities of producing X-class flares corresponded to sunspot classes F and E, penumbral class K and sunspot

distribution C. These results agreed with the fact that FKC and EKC are the sunspot groups that are most related to significant flares.

7.1.3 Research Resources

In this research, a wide range of experience was gained in order to make full use of the following resources:

- Many sources of solar data (e.g. NGDC, SWPC, and SOHO/LASCO) were studied and multiple programming languages (e.g. C++ and MATLAB) were used to implement the association algorithms and to integrate different applications under one computer platform.
- Many machine learning algorithms (e.g. SVMs, CCNNs, and RBFNs) were optimised and compared to study the associations.
- A statistical machine learning method (HMM) was used and its parameters were estimated using an iterative expectation maximization algorithm (Baum-Welch).
- Several computer tools were developed to process the solar data (catalogues and association datasets) and provide data in appropriate format for the use of MATLAB toolboxes and machine learning algorithms.

The work presented in this thesis can be developed further and the performances of some predictions are not as high as they might be because of some circumstances that will be addressed in the future. These circumstances are listed with suggested associated solutions in the next section.

7.2 Suggestions for Further Work

7.2.1 Integration with Other Technologies

One of the strengths of this work lies in its potential to be integrated with other technologies that are developed within the space weather research group at the University of Bradford. Examples for such integrations are listed as follows:

- As mentioned in Chapter 6 and shown in Figure 7.1, the HMM work can be integrated with the ASAP system (Colak and Qahwaji, 2009) using data fusion techniques to enhance the automated predictions generated by ASAP. So, ASAP detects and classifies sunspot groups by processing MDI images then it provides a real-time sunspot data. Sunspot properties for at least three previous days would be used by the HMM system that models the sunspot evolution patterns to provide predictions for the McIntosh classes and the sunspot areas for the next 24 hours. Also, matching historical evolution sequences can be provided by this subsystem for the detected active regions. ASAP's flare monitor predicts the possible flaring activity using its machine learning based rules. Hence, ASAP's flaring predictions could be improved using computerised models provided by the flare-sunspot HMM system as explained Chapter 6.

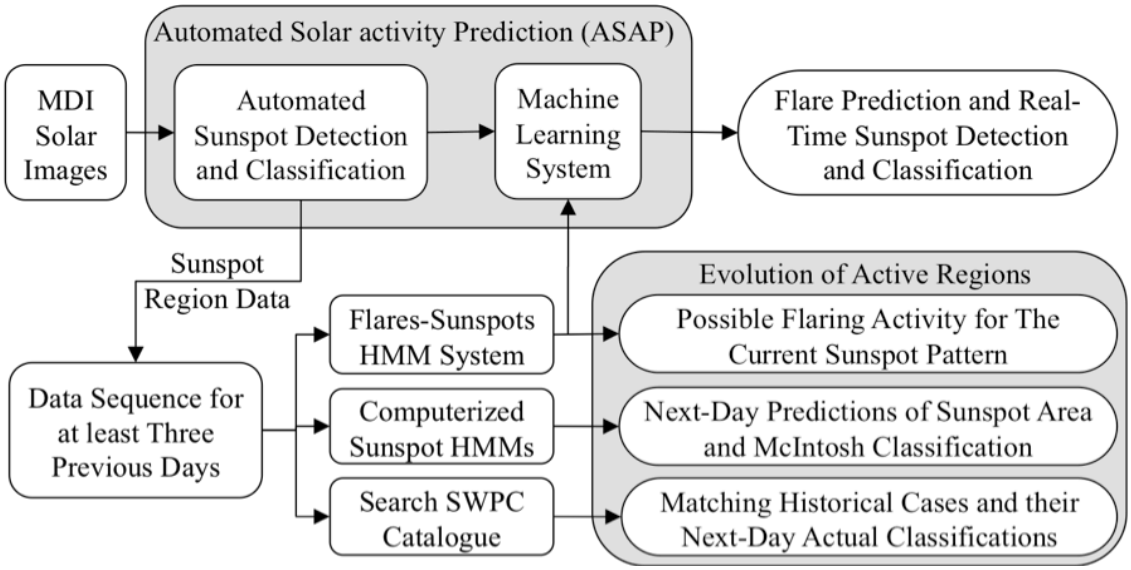


Figure 7.1 Future plan for inegrating the HMM work with ASAP.

- For a CME prediction system to be near real-time, real-time data needs to be inputted from other systems. For example, the filament detection and classification system included in Figure 7.2 is needed for the work of Section 5.5 to go online. It is planned to integrate the learning rules, related to CME predictions, with the ongoing research for the automated detection and

classification of filaments. In addition, the latest GOES X-ray flux data can be obtained from SWPC website[15]. Such integration would enable the use of the computerized rules for the purpose of an automated real-time CME predictions based on properties of the detected sunspots, filaments and flares.

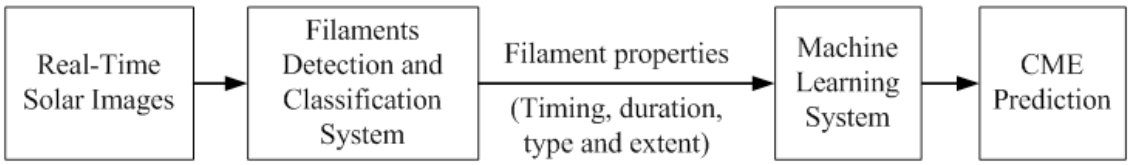


Figure 7.2 The hybrid CME prediction computer system based on CME-filament associations.

7.2.2 Improvements and Research Extensions

Some of the challenges that still need to be overcome, with suggested solutions and some ideas for further research are included in the following list:

- As concluded in Chapter 3, a large number of filaments are missing from the NGDC filaments catalogue. This clearly affected the findings in Chapter 5 as the data gaps in years 2000 and 2001 bias the SVM learning rules to predict incorrectly that filaments within this period are more likely not to initiate CMEs. To correct this bias it is necessary either to find another accurate filament catalogue or to create another more consistent one in future research.
- CMEs can be associated with erupting filaments/prominences and/or solar flares. However, in this thesis, CME associations with filaments and flares were considered separately using different association algorithms. To enhance the CME prediction accuracy it is necessary to combine both association algorithms in one platform, which would provide the ability to find the intersected data between these associations, hence enhancing the learning rules. This will be investigated in the near future.

[15] http://www.swpc.noaa.gov/rt_plots, last access: 2009.

- All the CME work reported in this thesis does not distinguish between front side and backside CMEs and it is possible for the present algorithms to associate a filament or a flare with a backside CME. For example, the CME-filament association algorithm associated a CME-filament pair on 30 Jun 1999 where the CME event was recorded at 13:31 and the filament was first observed at 12:55. However, it was reported in the preliminary list[16] of CME events, generated by the LASCO team, that this CME event was a partial halo backside event. The association algorithms have used most of the data reported in the catalogues without the use of solar images. There is only a small difference in the visibility of front-side and back-side CMEs, so it is very hard to distinguish them using only coronagraph observations (Yashiro et al., 2006). It is desirable to confirm that a CME originates from the front side by checking the lower corona images obtained by the Soft X-ray Telescope (SXT) on Yohkoh and the Extreme ultraviolet Imaging Telescope (EIT) on SOHO. This will be investigated in future work.
- It is believed that the predictions provided by the HMM system of Chapter 6 can be enhanced by a real time computer platform that searches the history for identical evolution sequences and suggests the possible next-day sunspot area and McIntosh class. The SWPC sunspots catalogue, used in the HMM work, consists of sunspot records that are taken at fixed times (one reading every 24 hours). Because of the consistency of the data, they were quite suitable for the study of Chapter 6 but it would have been generally more useful if more readings were taken per day. If the system performance reaches high accuracy values in the future, then the sunspot evolution patterns can be predicted while the active region under investigation is on the backside of the Sun.

[16] http://lasco-www.nrl.navy.mil/index.php?p=content/cmelist, last access: 2009.

References

AFWAMAN15-1 (2003) Sunspot Classification. *Air Force Weather Agency Manual.* Air Force E-Publishing.

AKIYAMA, S., GOPALSWAMY, N. & YASHIRO, S. (2006) The CME-Productivity Associated With Flares from Two ARs. *36th COSPAR Scientific Assembly.* Beijing, China.

AL-OMARI, M., QAHWAJI, R., COLAK, T. & IPSON, S. (2008) Support Vector Machines for Automated Knowledge Extraction from Historical Solar Data: A Practical Study on CME Predictions. IN SALEEM, A. I. & BARAKAT, S. (Eds.) *5th International Multi-Conference on Systems, Signals and Devices (IEEE SSD 2008).* Amman, Jordan.

AL-OMARI, M., QAHWAJI, R., COLAK, T. & IPSON, S. (2009a) Machine Learning-Based Investigation of the Associations between CMEs and Filaments. *SUBMITTED to Solar Phys.*

AL-OMARI, M., QAHWAJI, R., COLAK, T., IPSON, S. & BALCH, C. (2009b) Next-Day Prediction of Sunspots Area and McIntosh Classifications using Hidden Markov Models. *ACCEPTED in 2009 International Conference on CYBERWORLDS.* Bradford, UK.

ANDREWS, M. D. (2003) A Search for CMEs Associated with Big Flares. *Solar Physics,* 218**,** 261-279.

ASCHWANDEN, M. J. (2004) *Physics of the Solar Corona: An Introduction,* Chichester, UK, Praxis Publishing.

BAHL, L. R. & JELINEK, F. (1975) Decoding for Channels With Insertions, Deletions, and Substitutions With Applications to Speech Recognition. *IEEE Trans. Informat. Theory,* IT-21**,** 404-411.

BAKER, D. M. (1970) Flare Classification Based upon X-Ray Intensity. *AIAA Paper***,** 1370.

BAKER, J. K. (1975) The Dragon System - An Overview. *IEEE Trans. Acoust. Speech Signal Processing,* ASSP-23**,** 24-29.

BALCH, C. C. (2008) Updated verification of the Space Weather Prediction Center's solar energetic particle prediction model. *Space Weather,* 6**,** S01001.

BENGIO, Y. (1999) Markovian Models for Sequential Data. *Neural Computing Surveys,* 2**,** 129-162.

BERNASCONI, P., RUST, D. & HAKIM, D. (2005) Advanced Automated Solar Filament Detection And Characterization Code: Description, Performance, And Results. *Solar Physics,* 228**,** 97-117.

BORDA, R. A. F., MININNI, P. D., MANDRINI, C. H., GÓMEZ, D. O., BAUER, O. H. & ROVIRA, M. G. (2002) Automatic Solar Flare Detection Using Neural Network Techniques. *Solar Physics,* 206**,** 347-357.

BRIAND, C. (2003) Solar activity I: aspects of magnetic activity. *Astron. Nachr.,* 324**,** 357-361.

CARRINGTON, R. C. (1859) Description of a Singular Appearance seen in the Sun on September 1, 1859. *Monthly Notices of the Royal Astronomical Society,* 20**,** 13-15.

CLIVER, E. W. & HUDSON, H. S. (2002) CMEs: How do the puzzle pieces fit together? *J. Atmos. Sol. Terr. Phys.,* 64**,** 231-252.

COLAK, T. & QAHWAJI, R. (2007a) Automated McIntosh-Based Classification of Sunspot Groups Using MDI Images. *Solar Physics*.

COLAK, T. & QAHWAJI, R. (2007b) Automatic Sunspot Classification for Real-Time Forecasting of Solar Activities. *3rd International Conference on Recent Advances in Space Technologies, RAST '07.* Istanbul, IEEE.

COLAK, T. & QAHWAJI, R. (2007c) Hybrid Computer Platform for the Automated Prediction of Solar Flares. *SOHO 20 Conference: Transient Events on the Sun and in the Heliosphere.* Ghent, Belgium, http://www.soho20.org/IMG/pdf/soho20-Timetable-v8-final.pdf.

COLAK, T. & QAHWAJI, R. (2009) Automated Solar Activity Prediction: A hybrid computer platform using machine learning and solar imaging for automated prediction of solar flares. *Space Weather,* 7**,** S06001.

DAVIES, K. (1989) *Ionospheric Radio,* London, Institution of Electrical Engineers.

FAHLMANN, S. E. & LEBIERE, C. (1989) The cascade-correlation learning architecture. *Advances in Neural Information Processing Systems 2 (NIPS-2).* Denver, Colorado.

FAWCETT, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters,* 27**,** 861-874.

FREUND, Y. & SCHAPIRE, R. E. (1996) Game theory, on-line prediction and boosting. IN BLUM, A. & KEARNS, M. (Eds.) *Proc. Ninth Annual Conference on Computational Learning Theory.* Desenzano del Garda, Italy, ACM Press.

FREUND, Y. & SCHAPIRE, R. E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comp. System Sci.,* 55**,** 119–139.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2000) Additive logistic regression: A statistical view of boosting. *Ann. Stat.,* 38**,** 337–374.

FUKUNAGA, K. (1990) *Introduction to Statistical Pattern Recognition,* New York, Academic Press.

GELLERT, A. & VINTAN, L. (2006) Person Movement Prediction Using Hidden Markov Models. *Studies in Informatics and Control,* 15**,** 17.

GILBERT, H. R., HOLZER, T. E., BURKEPILE, J. T. & HUNDHAUSEN, A. J. (2000) Active and Eruptive Prominences and Their Relationship to Coronal Mass Ejections. *Astrophys. J.,* 537**,** 503-515.

GONZALEZ, W. D. & TSURUTANI, B. T. (1987) Criteria of Interplanetary Parameters Causing Intense Geomagnetic Storms (Dst< -100nt). *Planet. Space Sci.,* 35**,** 1101-1109.

GOPALSWAMY, N., AKIYAMA, S., YASHIRO, S. & MAKELA, P. (2009a) Coronal Mass Ejections from Sunspot and non-Sunspot Regions. IN HASAN, S. S. & RUTTEN, R. J. (Eds.) *Astrophys. Space Sci. Proc.* Heidelberg, Berlin, Springer-Verlag.

GOPALSWAMY, N., SHIMOJO, M., LU, W., YASHIRO, S., SHIBASAKI, K. & HOWARD, R. A. (2003) Prominence Eruptions and Coronal Mass Ejection: A Statistical Study Using Microwave Observations. *Astrophys. J.,* 586**,** 562-578.

GOPALSWAMY, N., YASHIRO, S., MICHALEK, G., STENBORG, G., VOURLIDAS, A., FREELAND, S. & HOWARD, R. (2009b) The SOHO/LASCO CME Catalog. *Earth, Moon, and Planets,* 104**,** 295-313.

GOSLING, J. T. (1995) Reply. *Journal of Geophysical Research,* 100**,** 7921–7923.

GREEN, L. M., HARRA, L. K., MATTHEWS, S. A. & CULHANE, J. L. (2001) Coronal mass ejections and their association to active region flaring. *Solar Physics,* 200**,** 189-202.

GREEN, L. M., MATTHEWS, S. A., VAN DRIEL-GESZTELYI, L., HARRA, L. K. & CULHANE, J. L. (2002) Multi-wavelength observations of an X-class flare without a coronal mass ejection. *Solar Physics,* 205**,** 325-339.

GUO, H. J. & CHAN, A. D. C. (2006) Approximated Mutual Information Training for Speech Recognition Using Myoelectric Signals. *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE.*

HALE, G. E., ELLERMAN, F., NICHOLSON, S. B. & JOY, A. H. (1919) The Magnetic Polarity of Sun-Spots. *Astrophys. J.,* 49**,** 153.

HATHAWAY, D., WILSON, R. & REICHMANN, E. (1994) The shape of the sunspot cycle. *Solar Physics,* 151**,** 177-190.

HEIDKE, P. (1926) Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst. *Geograf. Ann.,* 8**,** 301-349.

HILLARIS, A., PETOUSIS, V., MITSAKOU, E., VASSILIOU, C., MOUSSAS, X., POLYGIANNAKIS, J., PREKA-PAPADEMA, P., CAROUBALOS, C., ALISSANDRAKIS, C., TSITSIPIS, P., KONTOGEORGOS, A., BOUGERET, J.-L. & DUMAS, G. (2006) Solar Flares With and Without SOHO/LASCO Coronal Mass Ejections and Type II Shocks. *Advances in Space Research,* 38**,** 1007-1010.

HORI, K. & CULHANE, J. L. (2002) Trajectories of Microwave Prominence Eruptions *Astron. Astrophys.,* 382**,** 666-677.

JELINEK, F. (1969) A Fast Sequential Decoding Algorithm Using A Stack. *IBM J. Res. Develop.,* 13**,** 675-685.

JELINEK, F., BAHL, L. R. & MERCER, R. L. (1975) Design of A Linguistic Statistical Decoder For The Recognition Of Continuous Speech. *IEEE Trans. Informat. Theory,* IT-21**,** 250-256.

JING, J. (2005) Dynamics of Filaments, Flares and Coronal Mass Ejections (CMEs). *Federated Physics Department.* Newark, New Jersey, State University of New Jersey.

JING, J., YANG, G. & WANG, H. M. (2003) Statistical Studies of filament eruptions, flares and CME. *Bull. Am. Astron. Soc.,* 35**,** 815.

JING, J., YURCHYSHYN, V. B., YANG, G., XU, Y. & WANG, H. (2004) On the Relation between Filament Eruptions, Flares, and Coronal Mass Ejections. *Astrophys. J.,* 614**,** 1054-1062.

JONES, F. S. (1958) Classification of Solar Prominences. *J. R. Astron. Soc. Canada,* 52**,** 149-157.

JURAFSKY, D. & MARTIN, J. H. (2008) *Speech and Language Processing*, Prentice Hall.

KIEPENHEUER, K. O. (1953) Solar activity. IN KUIPER, G. P. (Ed. *The Sun.* Chicago, University of Chicago Press.

KLIMCHUK, J. A. (2001) Theory of Coronal Mass Ejections. IN SONG, P., SINGER, H. & SISCOE, G. (Eds.) *Space Weather, AGU Geophys. Monogr. 125.* Washington.

KOHLSCHEIN, C. (2006) An introduction to Hidden Markov Models. Probability and Randomization in Computer Science. Seminar in winter semester 2006/2007 at Aachen University.

KOSKINEN, H., TANSKANEN, E., PIRJOLA, R., PULKKINEN, A., DYER, C., RODGERS, D., CANNON, P., MANDEVILLE, J.-C. & D.BOSCHER (2001) Space Weather Effects Catalogue. *ESA Space Weather Programme Feasibility Studies.* FMI, QinetiQ, RAL Consortium.

KÜNZEL, H. (1960) Die Flare-Häufigkeit in Fleckengruppen unterschiedlicher Klasse und magnetischer Struktur. *Astronomische Nachrichten,* 285**,** 271.

KUROKAWA, H. (2002) Study of Energy Build-up in Solar Flares. *Journal of the Communications Research Laboratory. Special issue on Space Weather Forecast Study on Space Weather and its Hazards,* 49**,** 5-15.

LENZ, D. (2004) Understanding and Predicting Space Weather. *The Industrial Physicist,* 9**,** 18-21.

LIN, R. P. & HUDSON, H. S. (1976) Non-thermal processes in large solar flares. *Solar Physics,* 50**,** 153-178.

LIU, C., DENG, N., LIU, Y., FALCONER, D., GOODE, P. R., DENKER, C. & WANG, H. (2005) Rapid Change of δ Spot Structure Associated with Seven Major Flares. *Astrophys. J.,* 622**,** 722.

LOW, B. C. (1996) Solar activity and the corona. *Solar Phys.,* 167**,** 217-265.

LOW, B. C. (1999a) Coronal Mass Ejections, flares and prominences. IN HABBAL, S. R., ESSER, R., HOLLWEG, J. V. & ISENBERG, P. A. (Eds.) *Solar Wind Nine 471.* 1 ed. Nantucket, Massachusetts (USA), AIP.

LOW, B. C. (1999b) Magnetic Energy and Helicity in Open Systems. IN BROWN, M. R., CANFIELD, R. C. & PEVTSOV, A. A. (Eds.) *Magnetic Helicity in Space and Laboratory Plasmas.* Washington, AGU Geophys. Monogr. 111.

LOW, B. C. (2001a) Coronal mass ejections, magnetic flux ropes, and solar magnetism. *J. Geophys. Res.,* 106**,** 25141-25164.

LOW, B. C. (2001b) Solar Coronal Mass Ejection: Theory. IN MURDIN, P. (Ed.) *Encyclopedia of Astronomy and Astrophysics.* Bristol, Institute of Physics Publishing.

LOW, B. C., FONG, B. & FAN, Y. (2003) The Mass of a Solar Quiescent Prominence. *Astrophys. J.,* 594**,** 1060-1067.

MACQUEEN, R. M. & FISHER, R. R. (1983) The kinematics of solar inner coronal transients. *Solar Physics,* 89**,** 89-102.

MCINTOSH, P. S. (1990) The classification of sunspot groups. *Solar Physics,* 125**,** 251-267.

MENZEL, D. H. & EVANS, J. W. (1953) Convegno Volta. *Accademia Nazionale dei Lincei,* 11**,** 119.

MENZEL, D. H. & JONES, F. S. (1962) Solar Prominence Activity, 1944-1954. *J. R. Astron. Soc. Canada,* 53**,** 193-202.

MIKLOS, I. & MEYER, I. (2005) A linear memory algorithm for Baum-Welch training. *BMC Bioinformatics,* 6**,** 231.

MOON, Y. J., CHOE, G. S., WANG, H., PARK, Y. D., GOPALSWAMY, N., YANG, G. & YASHIRO, S. (2002) A Statistical Study of Two Classes of Coronal Mass Ejections. *Astrophys. J.,* 581**,** 694-702.

MOURADIAN, Z. (1998) The New "Solar Activity Synoptic Maps" of Observatoire de Paris. IN BALASUBRAMANIAM, K. S., HARVEY, J. W. & RABIN, D. M. (Eds.) *Synoptic Solar Physics - A.S.P.Conf. Ser. 140.*

MOURADIAN, Z., SORU-ESCAUT, I. & POJOGA, S. (1995) On the two classes of filament-prominence disappearance and their relation to coronal mass ejections. *Solar Physics,* 158**,** 269-281.

MUNRO, R. H., GOSLING, J. T., HILDNER, E., MACQUEEN, R. M., POLAND, A. I. & ROSS, C. L. (1979) The association of coronal mass ejection transients with other forms of solar activity. *Solar Phys.,* 61**,** 201-215.

OLIVER, D. W., KELLIHER, T. P. & KEEGAN, J. G. (1997) *Engineering Complex Systems with Models and Objects,* New York, McGraw-Hill.

PICK, M., LATHUILLERE, C. & LILENSTEN, J. (2001) Ground Based Measurements. *ESA Space Weather Programme Feasibility Studies.* Alcatel-LPCE Consortium.

POJOGA, S. & HUANG, T. S. (2003) On The Sudden Disappearances of Solar Filaments and Their Relationship with Coronal Mass Ejections. *Adv. Space Res.,* 32**,** 2641-2646.

POLAND, A. I., HOWARD, R. A., KOOMEN, M. J., MICHELS, D. J. & SHEELEY, N. R. (1981) Coronal transients near sunspot maximum. *Solar Phys.,* 69**,** 169-175.

QAHWAJI, R., AL-OMARI, M., COLAK, T. & IPSON, S. (2008a) Computerised Representation of the Association between Solar Features and Activities using Radial Basis Functions. IN VILLANUEVA, J. J. (Ed. *IASTED International Conference on Visualization, Imaging and Image Processing (VIIP 2008), ACTA Press.* Palma de Mallorca, Spain.

QAHWAJI, R., AL-OMARI, M., COLAK, T. & IPSON, S. (2008b) Using the Real, Gentle and Modest AdaBoost Learning Algorithms to Investigate the Computerised Associations between Coronal Mass Ejections and Filaments. IN MAHASNEH, J. (Ed. *Mosharaka International Conference on Communications, Computers and Applications (MIC-CCA 2008), Mosharaka for Researches and Studies.* Amman, Jordan, Mosharaka for Researches and Studies.

QAHWAJI, R. & COLAK, T. (2006a) Hybrid Imaging and Neural Networks Techniques for Processing Solar Images. *The International Journal of Computers and Their Applications,* 13**,** 9-16.

QAHWAJI, R. & COLAK, T. (2006b) Neural Network-based Prediction of Solar Activities. *3rd International Conference on Cybernetics and Information Technologies, Systems and Applications.* USA.

QAHWAJI, R. & COLAK, T. (2007) Automatic Short-Term Solar Flare Prediction Using Machine Learning and Sunspot Associations. *Solar Phys.,* 241**,** 195-211.

QAHWAJI, R., COLAK, T. & AL-OMARI, M. (2007a) Large-Scale Numerical Analysis for the Prediction of Flares using Support Vector Machines and Neural Networks. *SOHO 20 Conference: Transient Events on the Sun and in the Heliosphere.* Ghent, Belgium, http://www.soho20.org/IMG/Session4/SOHO20_S4-P60_qahwaji_Id=122.pdf.

QAHWAJI, R., COLAK, T., AL-OMARI, M., AHMED, O., ZRAQO, J. & IPSON, S. (2009) Present and Future Directions in the Automated Prediction of Solar Flares. *Space Weather Workshop: The meeting of Science, Research, Applications, Operations, and Users.* Boulder, Colorado,

http://www.fin.ucar.edu/UCARVSP/spaceweather/abstract_view.php?recid=1010.

QAHWAJI, R., COLAK, T., AL-OMARI, M. & IPSON, S. (2007b) Investigating the Association among Active Regions, Flares and CMEs using Machine Learning. *SOHO 20 Conference: Transient Events on the Sun and in the Heliosphere.* Ghent, Belgium, http://www.soho20.org/IMG/Session4/SOHO20_S4-P55_qahwaji_Id=120.pdf.

QAHWAJI, R., COLAK, T., AL-OMARI, M. & IPSON, S. (2008c) Automated Prediction of CMEs Using Machine Learning of CME-Flare Associations. *Solar Phys.,* 248**,** 471 - 483.

QU, M., SHIH, F., JING, J., WANG, H. & REES, D. (2004) Automatic Solar Flare Tracking. IN NEGOITA, M. G., HOWLETT, R. J. & JAIN, L. C. (Eds.) *Knowledge-Based Intelligent Information and Engineering Systems.* Berlin, Springer.

QU, M., SHIH, F. Y., JING, J. & WANG, H. (2003) Automatic Solar Flare Detection Using MLP, RBF, and SVM. *Solar Physics,* 217**,** 157-172.

RABINER, L. R. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE,* 77**,** 257-286.

ROBBRECHT, E., PATSOURAKOS, S. & VOURLIDAS, A. (2009) NO TRACE LEFT BEHIND: STEREO OBSERVATION OF A CORONAL MASS EJECTION WITHOUT LOW CORONAL SIGNATURES. *The Astrophysical Journal,* 701**,** 283-291.

RÜPING, S. (2000) mySVM-Manual University of Dortmund, Lehrstuhl Informatik 8.

S.E.LEVINSON, L.R.RABINER & M.M.SONDHI (1983) The Application of the Theory of Probabilistic Functions of A Markov Process to Automatic Speech Recognition. *The Bell System Technical Journal,* 62**,** 1035-1074.

SAKURAI, K. (1970) On the magnetic configuration of sunspot groups which produce solar proton flares. *Planetary and Space Science,* 18**,** 33-40.

SCHAPIRE, R. E. & SINGER, Y. (1999) Improved boosting algorithms using confidence-rated predictions. *Machine Learning,* 37**,** 297-336.

SCHOLL, I. & HABBAL, S. (2008) Automatic Detection and Classification of Coronal Holes and Filaments Based on EUV and Magnetogram Observations of the Solar Disk. *Solar Physics,* 248**,** 425-439.

SEVERNY, A. B. (1965) On the changes of magnetic fields connected with flares. *Stellar and Solar Magnetic Fields, Proceedings of the IAU Symposium no. 22. Edited by R. Lust. International Astronomical Union. Symposium no. 22, North-Holland Pub. Co.,.* Amsterdam.

SHEELEY, N. R., WALTERS, J. H., WANG, Y.-M. & HOWARD, R. A. (1999) Continuous Tracking of Coronal Outflows: Two Kinds of Coronal Mass Ejections. *J. Geophys. Res.,* 104**,** 24739-24768.

SHI, Z. & WANG, J. (1994) Delta-sunspots and X-class flares. *Solar Physics,* 149**,** 105-118.

SHRIVASTAVA, P. K. & SINGH, N. (2005) Latitudinal Distribution of Solar Flares and Their Association with Coronal Mass Ejections. *Chinese Journal of Astronomy and Astrophysics,* 5**,** 198–202.

SRIVASTAVA, N., GONZALEZ, W. D. & SAWANT, H. S. (1997) On the association of eruptive prominences, coronal holes and current sheets with the coronal mass ejections. *Adv. Space Res.,* 20**,** 2355-2358.

ST. CYR, O. C., BURKEPILE, J. T., HUNDHAUSEN, A. J. & LECINSKI, A. R. (1999) A Comparison of Ground-Based and Spacecraft Observations of Coronal Mass Ejections from 1980-1989. *J. Geophys. Res.,* 104**,** 12493-12506.

ST. CYR, O. C. & WEBB, D. F. (1991) Activity associated with coronal mass ejections at solar minimum: SMM observations from 1984–1986. *Solar Phys.,* 136**,** 379-394.

SUBRAMANIAN, P. & DERE, K. P. (2001) Source Regions of Coronal Mass Ejections. *Astrophys. J.,* 561**,** 372-395.

SUTTON, R. S. & BARTO, A. G. (1998) *Reinforcement Learning: An Introduction,* Cambridge, MA, MIT Press.

TASSOUL, J.-L., TASSOUL, M. & PHILLIPS, K. (2005) A Concise History of Solar and Stellar Physics. *Physics Today,* 58**,** 64-65.

TOUSEY, R. (1973) The Solar Corona. *Adv. Space Res.,* 13**,** 713.

VEZHNEVETS, A. & VEZHNEVETS, V. (2005) *Modest AdaBoost – teaching AdaBoost to generalize better,* Graphicon.

WARWICK, C. S. (1966) Sunspot Configurations and Proton Flares. *Astrophysical Journal,* 145**,** 215-223.

WEBB, D. F. (2000) Understanding CMEs and their source regions. *J. Atmos. Sol. Terr. Phys.,* 62**,** 1415-1426.

WEBB, D. F., CLIVER, E. W., GOPALSWAMY, N., HUDSON, H. S. & ST. CYR, O. C. (1998) The Solar Origin of the January 1997 Coronal Mass Ejection, Magnetic Cloud and Geomagnetic Storm. *Geophys. Res. Lett.,* 25**,** 2469-2472.

WEBB, D. F. & HUNDHAUSEN, A. J. (1987) Activity associated with the solar origin of coronal mass ejections. *Solar Phys.,* 108**,** 383-401.

WILSON, R. M. & HILDNER, E. (1984) Are interplanetary magnetic clouds manifestations of coronal transients at 1 AU? *Solar Phys.,* 91**,** 169-180.

YANG, G. & WANG, H. (2002) Statistical Studies of Filament Disappearances and CMEs. IN WANG, H. & XU, R. (Eds.) *Solar-Terrestrial Magnetic Activity and Space Environment, Proc. COSPAR Colloq. 14.* Beijing, China.

YASHIRO, S., GOPALSWAMY, N., AKIYAMA, S. & HOWARD, R. A. (2006) Associations of Coronal Mass Ejections as a function of X-ray Flare Properties. *36th COSPAR Scientific Assembly.* Beijing, China.

YASHIRO, S., GOPALSWAMY, N., AKIYAMA, S., MICHALEK, G. & HOWARD, R. A. (2005) Visibility of Coronal Mass Ejections as a Function of Flare Location and Intensity. *J. Geophys. Res.,* 110**,** A12S05.

YASHIRO, S., GOPALSWAMY, N., MICHALEK, G., ST.CYR, O. C., PLUNKETT, S. P., RICH, N. B. & HOWARD, R. A. (2004) A Catalog of White Light Coronal Mass Ejections Observed by the SOHO Spacecraft. *J. Geophys. Res.,* 109**,** A07105.

YEVLASHIN, L. S. & MALTSEV, Y. P. (2003) Relation between Coronal Mass Ejections, Solar Flares, Certain Parameters of the Magnetosphere, and Different Auroras during Great Magnetic Storms. *Geomagnetism and Aeronomy,* 43**,** 291–297.

YURCHYSHYN, V., WANG, H. & ABRAMENKO, V. (2003) How Directions and Helicity of Erupted Solar Magnetic Fields Define Geoeffectiveness of Coronal Mass Ejections. *Advances in Space Research,* 32**,** 1965-1970.

ZHANG, J., DERE, K. P., HOWARD, R. A., KUNDU, M. R. & WHITE, S. M. (2001) On the Temporal Relationship between Coronal Mass Ejections and Flares. *The Astrophysical Journal,* 559**,** 452-462.

ZHANG, J. & WANG, J. (2001) Filament Eruptions and Halo Coronal Mass Ejections. *The Astrophysical Journal,* 554**,** 474-487.

ZHANG, M. & LOW, B. C. (2004) Magnetic Energy Storage in the Two Hydromagnetic Types of Solar Prominences. *Astrophys. J.,* 600**,** 1043-1051.

ZHOU, G., WANG, J. & CAO, Z. (2003) Correlation Between Halo Coronal Mass Ejections and Solar Surface Activity. *Astron. Astrophys.,* 397**,** 1057-1067.

ZIRIN, H. & LIGGETT, M. A. (1982) Delta spots and great flares. *Solar Physics,* 113**,** 267-283.